EUROPEAN
INVESTMENT
FUND

EIF Research & Market Analysis

Working Paper **2016/38**

The European venture capital landscape: an EIF perspective

Volume II:
Growth patterns of EIF-backed startups

Simone Signore

Simone Signore is Research Officer in EIF's Research & Market Analysis team.

Contact: s.signore@eif.org

Tel.: +352 2485 81 636

Editor

Helmut Kraemer-Eis,

Head of EIF's Research & Market Analysis, Chief Economist

Contact:

European Investment Fund

37B, avenue J.F. Kennedy, L-2968 Luxembourg

Tel.: +352 248581 394

http://www.eif.org/news_centre/research/index.htm

Scan above to obtain a PDF version of this working paper

Luxembourg, December 2016

This page intentionally left blank

Abstract†

Start-up growth is often treated as a stylised fact, despite an extensive research body composed of divergent theories and empirical findings. Against this background, this work contributes to the literature by analysing a hand-collected dataset of 2,951 EIF-backed VC start-ups. Section 2 briefly reviews the relevant literature and 3 discusses data and methods. Section 4 uses descriptive statistics to explore average and median growth trends of start-ups, following an EIF-backed VC investment. Section 5 employs a *latent class cluster analysis* to establish a taxonomy of start-up growth profiles, characterised by speed and bias towards sales or patenting. The observed growth patterns, described in section 6, are typically idiosyncratic and persistent. However, a series of factors affecting growth mode can be evidenced. In particular, this paper finds that the geographic distribution of out-performing start-ups hints at national and regional factors acting as *enablers* for different typologies of successful growth. Implications for research and practice are discussed.

Keywords: EIF; start-up growth; venture capital; growth patterns

JEL codes: M13, G24, L25

† This paper benefited from the comments and inputs of many EIF colleagues. In particular, I would like to acknowledge the invaluable help of all my Research & Market Analysis colleagues. Moreover, I would like to express my gratitude for the helpful comments provided by prof. Thomas Hellmann, from Saïd Business School, and prof. Massimo Colombo, from Politecnico di Milano. All errors are attributable to the author.

Non-technical Summary

This work is the second volume of the series of working papers entitled "The European venture capital landscape: an EIF perspective". The series' goal is to quantify the economic effects brought by EIF-supported venture capital investments. Moreover, the series aims at assessing whether EIF's VC activity positively affected beneficiary start-up companies, contributing to the broader theme of government intervention in the field of venture capital.

This paper is mainly concerned with start-up growth, as measured through several economic and financial variables. The analysis consists of two separate but intertwined parts. The first block is dedicated to the analysis of start-up growth trends, *i.e.* the growth trajectories undertaken by start-ups in the aftermath of an EIF-backed VC investment. The analysis in section 4 leverages on data collected from Bureau Van Dijk's *Orbis* database, addressing missing data concerns through the use of a robust re-weighting methodology.

Using a wide range of descriptive statistics, section 4 documents the remarkable growth of EIF-backed start-ups, both on average and median terms. Average values of EIF-backed start-ups increase at least twofold for number of employees and total assets by the fourth year after investment date. Several profitability ratios indicate positive trajectories within a 7-year growth horizon. For instance, the proportion of firms with positive return-on-assets raises from 10% at investment date to 35%.

However, the insights evidenced by descriptive statistics are perturbed by the presence of extreme outliers and, in general, the high heterogeneity of growth trends. While outliers are a defining feature of the venture capital industry, their presence renders the identification of one "typical" growth profile a difficult task, since various may exist. Geographic location, main industry of activity and the year/period of investment are observed to influence average and median growth trends, but a significant degree of heterogeneity still persists.

For this reason, the second part of this work attempts at identifying the major *typologies* of start-up growth. Section 5 carries a cluster analysis that combines five different measures of firm medium-term development. Section 6 describes the four identified growth profiles: *a) under-performers*, representing almost 13% of the portfolio, *b) moderate performers*, constituting 55% of all investees, and two types of out-performers. These are *c) sale-based growers* and *d) patent-based growers*, representing 12% and 20% of the portfolio respectively.

Each of the identified profiles is characterised by the growth speed and/or bias towards sales or patenting. Under-performers experience mostly negative growth rates, bringing businesses on the brink of default. Moderate performers grow substantially in terms of economic size, but the growth is not followed by an increase of investment valuation, nor it is supported by significant patenting activity. Sale-based growers achieve an explosive 5-year growth driven by sales, while their investment valuation and patenting growth rates have lower levels. Patent-based growers show the highest patenting growth rates, as well as the highest valuation growth rate.

Growth profiles tend to be persistent over time: in most cases, it is more likely that start-ups hold on to their profile than transition to another. If convergence towards a certain state is observed, then it typically leads to more moderate growth. However, the medium-term horizon on which growth types are based is shown to bear low predictive ability towards investments returns. While higher-growth profiles are certainly linked with return premia, all profiles still face a substantial risk of investment write-off, which is at best in the range of 30%. This is primarily due to the aftermath of the *dot-com* crash: recently, high-growth start-ups significantly outperformed under- and moderate- performers in terms of exit class. The second part of this paper concludes with an exercise in extrapolation, which documents the existence of numerous geographical clusters where one of the two out-performing profiles tends to prevail on the other.

This work presents numerous policy implications: on the one hand, it discusses the defining traits that compose the "genetic code" of EIF-backed — and possibly, *non* EIF-backed — European start-ups. On the other hand, it acknowledges the heterogeneity of start-up growth trajectories and attempts at identifying a number of profiles of growth. While the analysis concludes that growth trajectories are typically idiosyncratic and persistent, some determinants of growth mode are evidenced (*e.g* age, sector). Finally, the geographic distribution of out-performing start-ups hints at the presence of national and/or regional factors that may act as *enablers* of particular types of successful growth. Overall, the findings highlight the potential for EIF-backed VC start-ups to significantly contribute to the economic development and job creation across several regions of Europe.

Table of Contents

List of Figures

1 Introduction

The European venture capital landscape: an EIF perspective is a series of working papers redacted by the Research & Market Analysis team. The series' goal is twofold: on the one hand, the quantification of the economic effects brought by EIF-supported venture capital investments. On the other hand, the initiative aims at assessing whether EIF's VC activity positively affected beneficiary start-up companies,[1] contributing to the broader discussion on whether government intervention in the field of venture capital is both effective and economically justified.

Kraemer-Eis *et al.* (2016) described the overarching economic rationale of EIF's activities in the EU VC market, as well as the role the institution plays within the broader European VC ecosystem. By design, the entirety of EIF-supported venture capital investments has been the subject of analysis in the series' opener. As opposed to the part, the *whole* suited best an introductory exposition to EIF's venture capital activity.

Building on such introductory work, the purpose of this current issue — shared by most other forthcoming works in the series — is to depart from *macro*-related aspects and delve into the details of our company-level dataset of EIF-backed VC startups. To meet such goal the series will proceed stepwise, addressing policy-relevant aspects related to key actors in the venture capital ecosystem.

The first subject of analysis is the economic and financial growth of startups ensuing an EIF-backed VC investment. Following a quick introduction to the relevant theoretical and empirical research, this paper's first goal is descriptive in nature: to quantify startups' growth as measured through several economic and financial variables. The analysis will quickly touch on factors that are known to significantly affect growth trends (*e.g.* geographic location, main industry of activity and the year/period of investment). The analytical part of this work centres on a cluster analysis, aimed at establishing a taxonomy of VC-backed start-ups based on their growth trajectory. This paper concludes with a geographic analysis, focusing on regions of Europe where certain growth profiles seem dominant.

Despite its intuitive and broad notion, it is important to clarify the concept of "growth" analysed in this study: first, growth indicates here the entire firm development process, which can be characterised by expansion as well as decline. Second, while studies so far mostly focused on firm size, this paper also tackles the development of profits and the start-up's financial structure. Last, this work will be mainly concerned with *organic growth*, excluding growth generated via external acquisitions. Although insights on this second type of growth can be considered equally relevant, preliminary analysis shows that cases of acquisitive growth for EIF-backed VC start-ups are exceptionally rare.

2 A brief review of firm growth literature

The topic of business growth, and new ventures growth in particular, transcends the study of VC-backed companies, and as such benefits from a wealth of economic research. McKelvie and Wiklund (2010) identify three main streams of research that focus on firm growth, which can be conveniently simplified into *pre-*, *post-* and *mid-growth*. The first, labelled *growth as an outcome*, is aimed at

[1] Throughout this paper, the terms *startup*, *start-up* and *start-up company* will be used interchangeably.

identifying the determinants that lead to firm growth. The authors refer to this stream as the most popular, albeit equally inconclusive in its effort to isolate robust predictors of firm growth. The second stream is more concerned with the organisational changes in the aftermath of company growth, while the third and last focuses on the growth process itself, analysing firm organisational development while it experiences growth.

Although the authors emphasize the abstract nature of such classification, McKelvie and Wiklund (2010) conclude that the integration of these viewpoints, and a particular accent on the growth process, is crucial to stimulate further progress in this field of research. Their view is shared by Gilbert *et al.* (2006), whose review of the main determinants of firms' growth laments a complete lack of consideration for the "*how and where new ventures grow*" (*ibid,* p. 945). The authors argue that these two key decisions affecting new ventures' growth — whether to have an organic or acquisitive growth (the *how*), and a domestic or international development (the *where*) — are pivotal points in the ascent of successful firms.

Consolidating an extensive, convoluted and multi-disciplinary literature, Coad (2007) provides perhaps the most comprehensive review of firm growth research to date.[2] Coad's work provides both a review of the major determinants of firm's development (*e.g.* age, innovation, industry, geography), as well as a survey on a wide range of theories on firm growth. Notwithstanding, the main take away from the literature on firm growth seems to be the recognition that significant heterogeneity in growth patterns exists, both in studies that compare *small* firms against *large* as well as in those exclusively focusing on small firms.

As such, a first hypothesis can be formulated: growth trends of EIF-backed start-ups are expected to show high heterogeneity. This proposition is verified in section 4. However relevant, it is difficult to think that such first hypothesis could provide significant policy insights. Indeed, the identification of "stable" clusters containing start-ups with similar growth profiles would prove more valuable, as it would help to establish a *taxonomy* of start-ups according to their growth trends. To address this relevant objective, a second hypothesis complements the first: despite the significant heterogeneity, a number of start-up growth profiles can be isolated among EIF-backed start-ups. To verify such claim, section 5 and section 6 discuss the application of a *latent class cluster analysis*.

This paper is not the first to employ cluster analysis to explore the heterogeneity of start-up growth: Delmar *et al.* (2003) perform an exploratory cluster analysis of 1,501 Swedish new high-growth ventures founded in the 1987-1996 period. The authors find up to seven distinct growth patterns. Differences among growth clusters were determined by the bias towards a specific growth indicator (*e.g.* sales, organic or overall employment), whether the growth was absolute (*i.e.* significant also in absolute terms) or relative (*i.e.* due to base effects), and whether growth was steady or erratic. While Delmar *et al.* (2003) exclusively focus on high-growth new ventures, this paper contributes to the literature by including growth patterns of non-successful and moderately successful start-ups.

Closing the circle, the identification of growth profiles could play a key role in addressing the theme of growth determinants. Within the existing literature, a compelling body of research has tried to motivate differences in growth trajectories by means of a "*stages of growth*" theory (Kazanjian and

[2] Substantial parts of Coad's preliminary survey were subsequently published in Coad (2009).

Drazin, 1990), where new ventures evolve through a series of sequential stages which pose different challenges (but also opportunities) to further growth. *Stages of growth* models tend to focus on — and better predict — the growth of technology-based start-ups, on which substantial empirical evidence has been produced in the last decades. For instance, Almus and Nerlinger (1999) show that innovative start-ups grow faster on average than non-innovative firms.[3]

As technology-based start-ups constitute the typical target of venture capital investors, many studies have also focused on the impact of VC on firm growth. Among these, Davila *et al.* (2003) use data on 494 VC- and non-VC-backed start-ups in the San Francisco bay area, observing that the presence of venture capital is related to faster employee growth. Puri and Zarutskie (2012) compare VC-backed against non-VC-backed US start-ups exploiting 25 years of longitudinal data, observing that VC investments positively affected, among others, employment and size growth. Finally, Bertoni *et al.* (2011) and Bronzini *et al.* (2015) use two different hand-collected datasets of Italian start-ups to find that VC-backed companies experience a growth premium in multiple size and profit indicators.

Thus, the role of VC investments should not be underestimated when analysing growth trends of invested companies. This calls for a last crucial hypothesis, whose assessment lies beyond the scope of this work (but within the scope of the overall series). Namely, that growth trends of VC-backed companies supported by EIF significantly differ from those of start-ups not backed by VC investments.

3 Data and methods

The analysed EIF VC portfolio consists of a hand-collected dataset of 2,951 seed and start-up stage companies,[4] supported by one or more EIF-backed VC funds in the 1996–2014 period. This dataset contains only EIF-backed VC startups whose size, age and industry comply with the canons of conventional VC-targeted companies. In the final collection step, startups' identities were linked to their financial profiles, provided in Bureau Van Dijk's *Orbis* database.[5]

Building on EIF's fund-of-funds approach, data on EIF-backed startups can be described by exploiting the analogies it bears towards conventional portfolios of startups. Specifically, one can use year-by-year aggregates of actively invested companies to depict the dynamic trends of the EIF VC portfolio. "Actively invested" companies have at least one EIF-backed fund as shareholder.

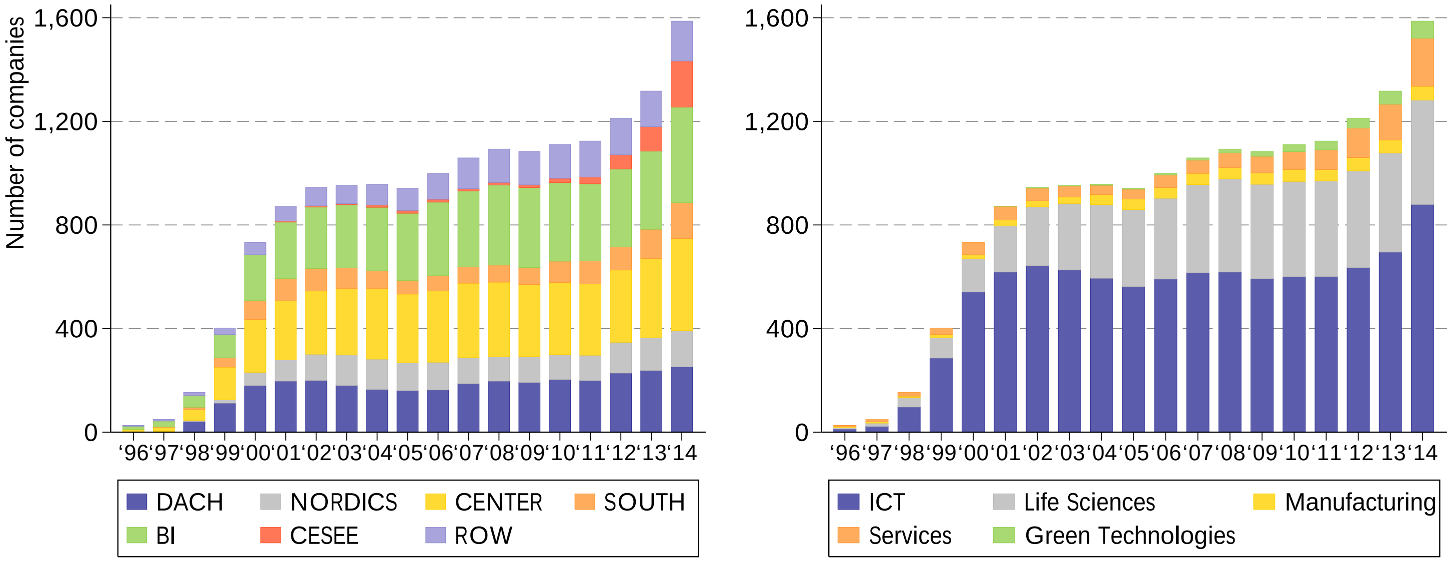
Figure 1 splits the dynamics of the EIF VC portfolio by macro-region and macro-industry. Figure 1a shows that the bulk of the EIF-backed startups are located in three key areas: DACH, CENTER (France and Benelux) and the British Isles (UK and Ireland). Moreover, there is a visible trend towards the geographical balancing of the EIF VC portfolio, evidenced by the increasing shares of Nordics, Southern European and CESEE investees. This confirms the previous findings in Kraemer-Eis *et al.* 2016, which provides an extensive review of the geographical features of EIF-backed startups.

[3] Their work also uses employment growth to indicate firm development.

[4] Since the first publication, mentioning 2,934 companies, the dataset was augmented with a small number of additional investments. The restricted number of additions caused no tangible change to prior findings.

[5] Bureau Van Dijk's (BvD) Orbis is a database containing information on over 200 million enterprises active worldwide (as of December 2016). Orbis contains up to 51 different firm-level financial indicators, either sourced from the firm's balance sheet or P&L account, and a series of 26 computed ratios. Recent additions to the database also include the number of patents and trademarks.

Figure 1: Key features of the EIF VC portfolio[6]



(a) Actively invested VC startups by geographical macro-area[7]

(b) Actively invested VC startups by main industry of activity

Note: each bar aggregates active portfolio companies in the reference year. Hence, the chart is not cumulative, as exited companies will drop out of the sample and not accounted for in subsequent years.

Repeating the exercise, Figure 1b shows portfolio trend across five main industries, or "macro-industries", resulting from the aggregation of 16 different types of industrial activity.[8] Figure 1b evidences how EIF-backed VC financing has predominantly targeted ICT companies, while at the same time it displays how an increasing proportion of investments has reached out to life science, green-tech and service-based companies. At the same time, it is worth mentioning that the 2013-14 biennium saw an uptake of the number of supported ICT startups, almost exclusively driven by a rise in the number of software-related firms.

As reported in Kraemer-Eis *et al.* (2016), more than 91% of all start-ups discussed above have been linked to their respective entries in the Orbis database. However, the *matching* of a company's profile does not imply that its data is readily *usable* to generate descriptive statistics. In fact, a considerable share of information on matched companies is unobservable: while the precise percentage of usable companies will depend on the specific economic or financial indicator under scrutiny (see Figure 2), overall the coverage rate of the EIF VC portfolio ranges from 25% to 40%.

The consequences of missing data lie not so much in the rate of missings, which *ceteris paribus* merely causes increased uncertainty (*i.e.* higher estimate variance), but in the *pattern* of missing data. An important methodological contribution of this work lies in the use of *stratified response propensity weights* (Little, 1986) for the estimation of portfolio statistics from Orbis data.

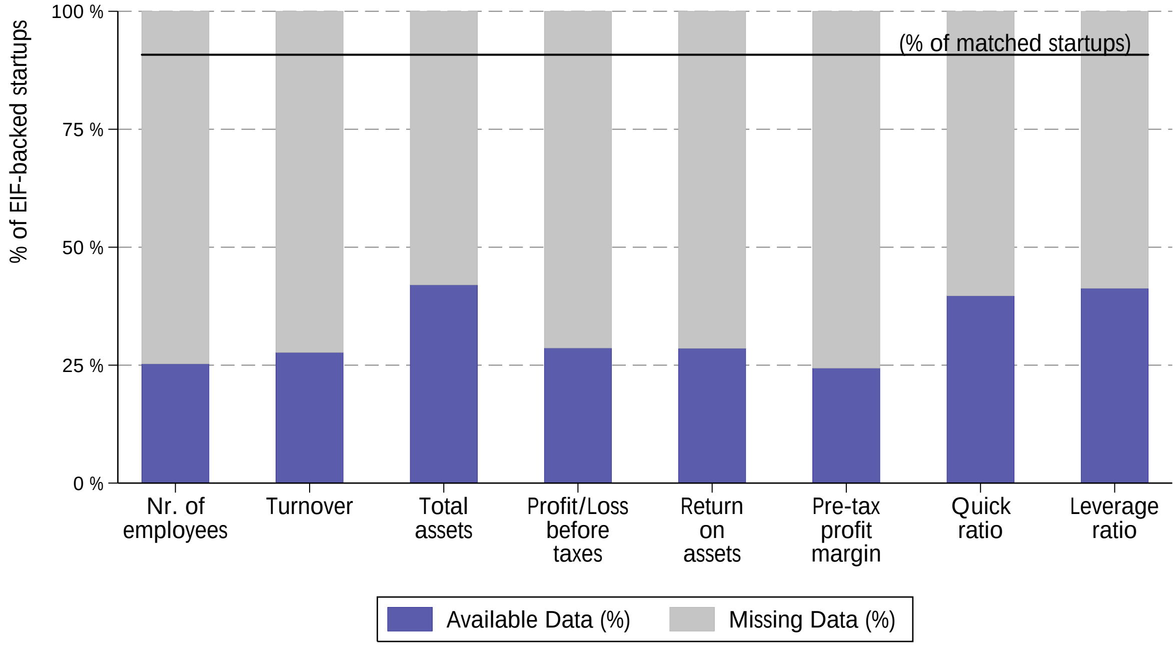
---

[7] Unless otherwise stated, all figures in this research are an elaboration of the author, based on EIF data.

[7] DACH: AT, CH, DE; NORDICS: DK, FI, NO, SE; CENTER: BE, FR, LU, NL; SOUTH: GR, ES, IT, MT, PT; BI (British Isles): IE, UK; CESEE: BG, CZ, EE, LT, LV, PL, RO, SK, TR, CY; ROW (Rest Of the World): AR, AU, CA, CN, CR, HK, IL, IN, MX, PH, RU, SG, US, UY

[8] The sectoral nomenclature follows Invest Europe's industry classification. Additional details on startup's sectors and their relationship to NACE Rev. 2 classification codes is included in Appendix A.

Figure 2: Incidence of missing values (portfolio coverage per financial indicator)



Note: Details on the methodology used to derive missingness rates are included in Appendix B. Appendix C provides definitions for each of the indicators portrayed above.

The technique bears a certain analogy with response adjustments in surveys.[9] Indeed, by performing an exercise in abstraction, it could be argued that data collected from Orbis mirrors the process of gathering data directly from companies through a well-designed survey: companies are first contacted (*matched*); a number of performance-related answers is then collected (*usable*). Exploiting the features of the data-collection process and several auxiliary variables, it becomes possible to address the bias introduced by missing data. The approach is discussed in detail in Appendix B.

4 Descriptive analysis

A descriptive analysis is carried using 14 distinct economic and financial indicators. The indicators follow three broad concepts of firm development: *size*, *profitability* and *financial structure*.[10]

To appreciate firm development and combine data on start-ups invested over a time window of 20 years, trends are examined since the first year the company benefits from an EIF-backed investment ("year zero"). If data allows, all firms are then followed up to the seventh year post-investment.

The virtue of this approach is accompanied by a few limitations. First, nominal monetary values across such wide time period would not be comparable: to address this, all balance sheet monetary values are deflated with sector-based producer price indices and presented in constant EUR prices, with base year 2005.[11] Second, presenting yearly data in such terms introduces selection

[9] Despite a number of recent unfortunate applications, sampling theory in the context of surveys benefits from a long-established body of authoritative research.

[10] Appendix C provides a summary table of the financial indicators used in the analysis.

[11] Yearly producer price indices (PPIs) are assigned according to the country and sector of the start-up. Refer to Appendix A for a correspondence between EIF sectoral nomenclature and NACE Rev. 2 classification.

bias, which becomes more significant the farther the measurement year is from the first investment year. To give an example, consider that the last observed investment cohort is 2014, and that most recent financial data may be available for year 2016: this implies that observing the statistic on firm performance ”3 years after first investment date” automatically excludes the 2014 cohort. Although one straightforward approach to limit selection bias is to choose a ”growth horizon” composed of a relatively small number of years (this strategy is pursued here by examining trends at most seven years after the first investment), it is not possible to counter this effect when visualising the entire EIF VC portfolio trends. Alternatively, one could isolate and contain this bias by grouping adjacent cohorts and separately examine their trends. The results of this latter strategy are discussed in Appendix D.

Last, even if the analysis were to be carried on single cohorts,[12] trends would still be affected by survivorship bias. Survivorship bias, a specific case of selection bias, occurs when companies with a specific set of characteristics are disproportionately more likely to default, hence to be excluded from further performance estimates. Extensive research literature shows how smaller firms tend to be riskier, more likely to default. In such case, one can expect active companies in later post-investment years to be disproportionately larger in size, causing inflated average and median estimates.

Survivorship bias alters the interpretation of the estimated growth trends: estimates computed at any given post-investment year exclusively concern the subset of firms that survived until such year, thereby reducing the reliability of comparisons across subsequent post-investment periods. A pragmatic approach to this shortcoming would suggest yet again to focus on a small number of post-investment periods. In the EIF VC portfolio, the observed survival rate at 3, 5 and 7 years after investment is respectively 97%, 92% and 83%. While the attrition rate after 7 years raises some rightful concerns on the validity of the observed statistics, performance rates until such time can be considered comparable. Nevertheless, survivorship bias remains a pervasive distortion that particularly affects the analysis of young, high-risk companies. As such, forthcoming works in this Working Paper series will focus on this specific issue so as to provide a comprehensive overview of the phenomenon. To summarise the effects of missing data and sample selection, Appendix D reports sample sizes for each period and indicator used in the paper.

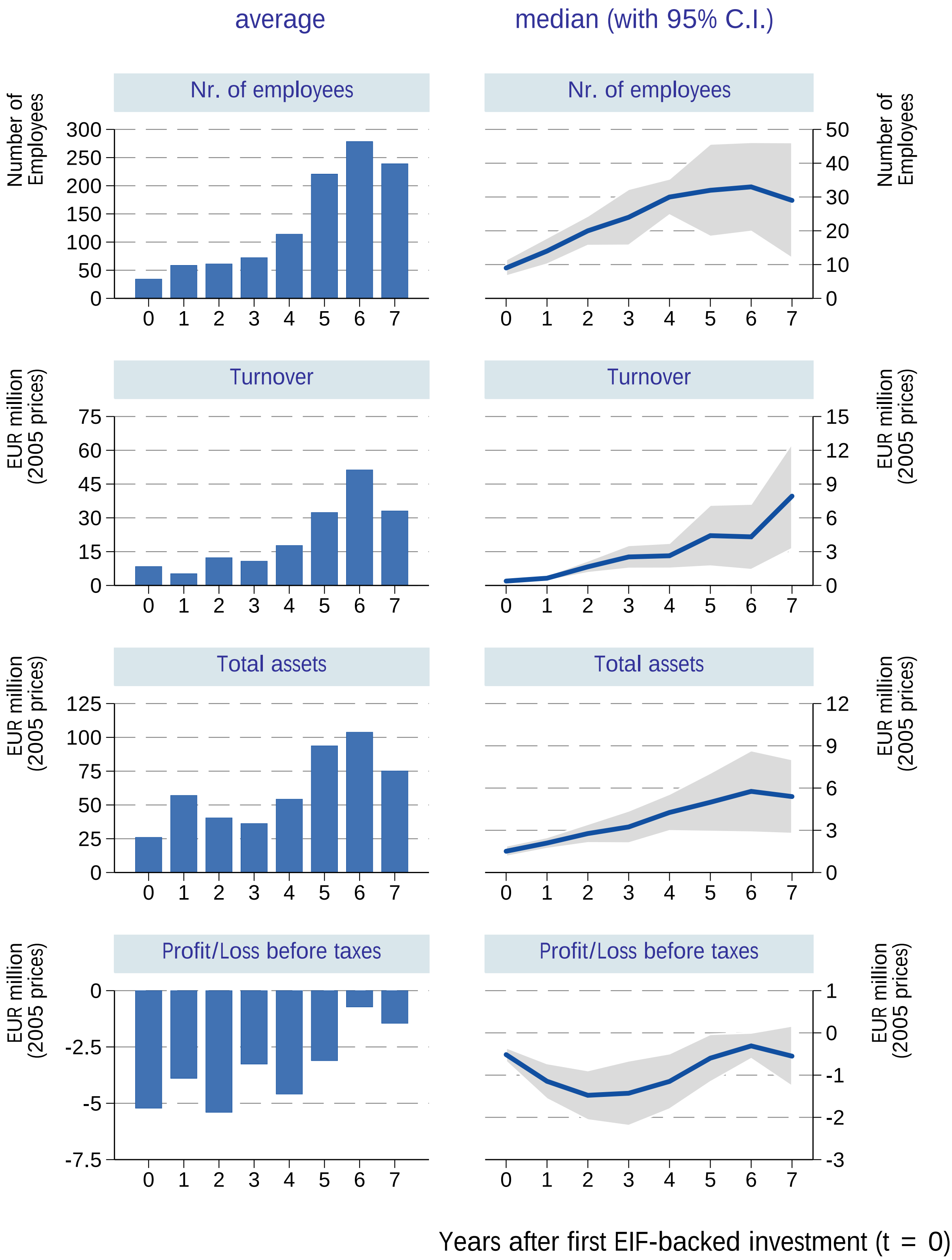
4.1 Economic size of start-ups

Size indicators consist of financial variables that typically define business size (number of employees, turnover and total assets), and are often utilised in the economic literature to represent firm growth. To these, this paper adds information on pre-tax profit and loss (P&L). Despite P&L being often assigned to the realm of financial performance, its trend offers valuable insights towards start-ups' overall growth. Both average and median estimates are computed in order to highlight general features of the underlying distribution. Median estimates are enriched with confidence bands indicating the degree of uncertainty of the estimate.[13] The evolution of start-up size indicators is portrayed in Figure 3, where average estimates are shown on the left and median estimates and their confidence intervals are portrayed on the right.

[12] Unfortunately this strategy cannot be pursued, given that data limitations would cause the occurrence of very small subsets of data with no representative power.

[13] Intervals at 95% confidence level for the median are estimated following the approach in Woodruff (1952), using jackknife standard errors. Confidence intervals for the mean omitted to facilitate the exposition.

Figure 3: Average and median growth trends of size indicators



Note: the figure above portrays size levels following an EIF-backed VC investment. Left-side charts show average values, while right-side charts show median values. The *x-axis* counts the periods (in years) following the VC investments, where period 0 is the investment year. Statistics are computed using *response propensity weights* (Little, 1986). Methodological details are discussed in Appendix B. All monetary values expressed in constant EUR prices, with 2005 as the base year.

There are three major features evidenced by Figure 3: first, subject to all potential biases expressed in the previous paragraph, both average and median trends indicate significant size growth for a substantial share of EIF-backed start-ups. For instance, the median employment and asset size four years after investment is about three times bigger than at investment date, and persists thereafter.

Second, the mean consistently exceeds the median, indicating a right skewness of the underlying distribution. Medians should thus be perceived as more reliable, representative measures of the "typical" evolution of EIF-backed start-ups. The heterogeneity of growth trajectories is explored in Appendix D, where the influence of the growth determinants discussed by Coad (2007) is tested by means of further descriptive statistics.

Third and last, profit trends visibly follow a J-shaped evolution, although there is a clear setback in the seventh year after investments, both for average and median values. The tapering of trends in this period is shared by all indicators of Figure 3. This seems to be caused by specific subgroups (*e.g.* ICT start-ups in the British Isles for the case of employees), as evidenced in Appendix D. Selection bias may be another major driver of the observed pattern.

4.2 Start-up profitability and financial structure

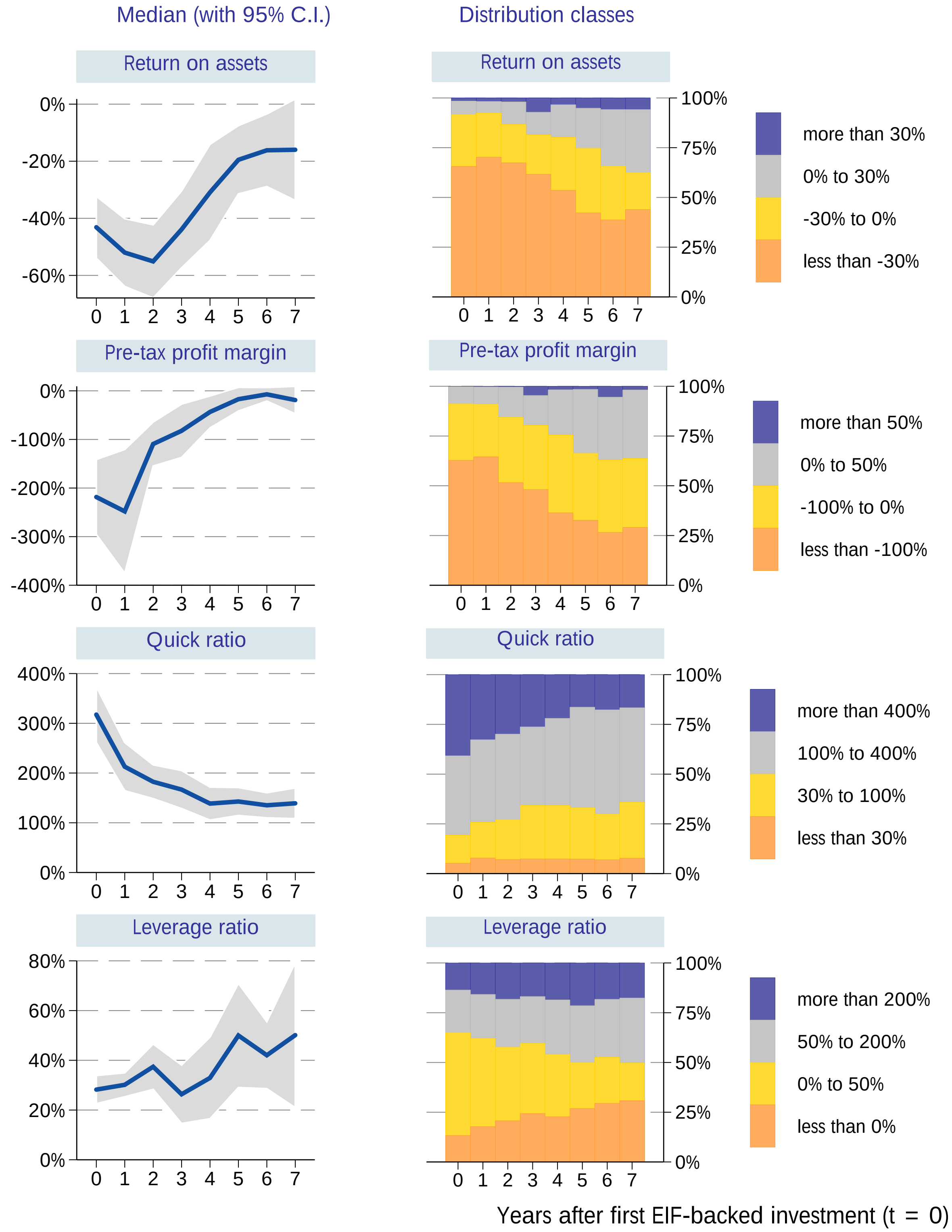
Profitability indicators are a direct derivation of the indicators explored in the previous section: return-on-assets (ROA, *i.e.* pre-tax P&L over total assets) and pre-tax profit margin (*i.e.* pre-tax P&L over turnover). The financial structure of start-ups is assessed via two additional ratios: the quick ratio (defined as current assets, net of inventories, divided by current liabilities) and the leverage ratio (*i.e.* total liabilities over total equity).

Return-on-assets and pre-tax profit margin are two widespread measures of firm profit efficiency: the first compares profits to firm overall value, while the second to firm revenues. The quick ratio is a classical measure of company liquidity, explicative of the firm's ability to repay its short term liabilities. The leverage ratio allows to glance through the company's level of indebtedness. Compared to size indicators, the analysis of these ratios further exposes the heterogeneity of the underlying distribution, as well as the presence of extreme outliers. For this reason, statistics for these variables are reported by focusing on median values (and their 95% confidence bands) as well as distribution classes arbitrarily defined. The results of this analysis are portrayed in Figure 4.

Figure 4 should not surprise readers experienced with start-up economics. Start-up companies in their year of investment are typically not generating profits (hence facing negative ROA and profit margin), highly liquid — a direct consequence of the EIF-backed VC investment — and with considerably low leverage ratio brought by their notorious difficulty to attract debt financing. However, already within a seven-year investment period it is possible to observe how for the majority of companies these ratios appear to converge to sustainable values.

While J-shaped trends are evidenced by both ROA and profit margin, it is never the case that these two ratios are mostly positive in the first seven years after investment date. However debatable the extent to which start-ups' long-term profitability truly affects investors decisions, this finding hints at one of the distinguishing features of the venture capital market, *i.e.* profit realisation for start-ups typically requires a long-term perspective.

Figure 4: Average and median growth trends of profitability and financial structure indicators



Note: the figure above portrays profitability and financial structure trends following an EIF-backed VC investment. Left-side charts show median values, while right-side charts show the proportion of companies populating each of the defined classes. The *x-axis* counts the periods (in years) following the VC investments, where period 0 is the investment year. Statistics are computed using *response propensity weights* (Little, 1986). Methodological notes discussed in Appendix B.

Appendix E complements the descriptive analysis by inspecting an additional set of start-up indicators. These share a similar nature with the financial ratios presented above, but their analysis further evidences the insights discussed so far.

5 Cluster analysis

The goal of cluster analysis is to explore data and assess whether (or not) it can be meaningfully characterised and summarised by a small number of groups, *i.e.* subsets of the original population. The groups, referred to as *clusters*, are formed in such a way that observations within groups tend to resemble each other, while observations between clusters tend to differ significantly. Moreover, cluster analysis is a viable *data reduction* process for the analysis of multi-dimensional phenomena, making this technique particularly useful in the analysis of complex processes such as firm growth.

In the terminology of Everitt (2011), there are three major approaches to cluster analysis: first, *hierarchical clustering* typically starts treating each observation as a single cluster, then proceeds to combine groupings based on a similarity measure. Second, *optimisation clustering* consists of a series of algorithms which seek to create a pre-determined number of clusters via minimisation or maximisation of a numerical criterion (usually based on similarity distances). The first two approaches are inherently heuristic, causing different implementations to yield results that are not comparable. For this reason, cluster analysis is often regarded as a semi-objective quantitative approach, for a lack of formal rules to model selection (*e.g.* the choice of a number of clusters, the choice of a distance measure).

An additional emerging approach towards cluster analysis employs a formal statistical model which assumes that the observed data is the results of a finite number of *clusters*, each characterised by a different multivariate distribution of the clustering variables. In such framework, the population distribution becomes a *finite mixture density* — a distribution resulting from the combination of other distributions — and the researcher can use appropriate models to estimate the parameters of such distribution.[14] By virtue of its parametric approach, *model-based* clustering does not simply predict distribution classes, but *probabilities* of being affiliated to a given group. Model-based clustering bears significant advantages, in that it offers a way to objectively compare different model specifications (*e.g.* by looking at different measures of model selection criteria). Moreover, model-based cluster analysis enables the clustering of structured data (*e.g.* panel data with repeated observations per individual). For instance, the statistical model can be tailored to account for time dependency in panel data, or alternatively it allows the researcher to set-up a model for a specific time horizon, then analyse *cluster transition* in subsequent periods.

Overall, model-based cluster analysis is considered superior to other heuristic approaches, with the only shortcoming of requiring sample sizes large enough in order to obtain reasonably precise parameter estimates (Everitt, 2011, p. 186). Against such background, this work discusses the results of a model-based cluster analysis of start-up growth. To ensure the robustness of the approach, the results are compared to both hierarchical cluster analysis (performed via Ward's method and

[14] Skrondal and Rabe-Hesketh (2004) note that finite mixture modelling can be seen as a form of *latent variable analysis*. Consequently, this approach has also been labelled *latent class cluster analysis*.

Euclidean distances) and *k-means* cluster analysis on Euclidean distances. Appendix F provides a detailed discussion of the methodological steps undertaken in the remainder of this chapter.

Start-up growth is assumed to be five-dimensional, *i.e.* composed by five different measures. Growth determinants are mostly based on the size measures used in section 4: number of employees, turnover, total assets. To these, the valuation of the company (derived from VC investor reports) and the number of patents are added to attain a more comprehensive notion of start-up growth.[15] The inclusion of company valuation – more precisely, the impossibility to observe this value once the start-up investment reaches the exit stage – also shapes the interpretation of the growth horizons: start-up growth will be measured as long as it is supported by the EIF VC activity.

As this research is concerned with growth patterns, each of the five variables is used to generate compound annual growth rates (CAGR) at different time horizons.[16] The choice of the time horizon is once again not trivial: as described in section 4, there are numerous sources of bias that affect the reliability of results when analysing long-term growth horizons. On the other hand, short-lived growth horizons also pose an accuracy issue, *i.e.* that such growth trends may not accurately predict the company's long- term trends. Against this background, the analysis in this chapter is carried by focusing on a 5-year growth horizon, which on one hand benefits from a relatively low level of sample attrition (see section 4), and on the other it is approximately equal to the median and average holding period of EIF-backed investments. For completeness, section 6.3 uses additional growth horizons and compares results in order to assess the persistence of growth profiles.

As opposed to sections 4.1 and 4.2, the calculation of CAGRs can only be performed for start-ups that have faced a wide-enough growth horizon: in other words, a 5-year growth horizon automatically excludes investment vintages beyond 2010.[17] While this is an additional source of data loss, it also provides a mean to compare growth rates in a uniform way. Data loss for this analysis is significant. Because of such "observed-horizon" requirement, one quarter of all portfolio companies are dropped. Moreover, start-ups located in "Rest of the World", Luxembourg and Greece had to be dropped due to almost-complete lack of growth data. All in all, about 65% of the original portfolio companies can be used in the analysis. Of these, more than 20% have non-missing growth data: the methods discussed in section 3 are thus used to address the bias brought by missing data.

In conclusion, 5-year CAGR of the above-mentioned growth factors are employed in the model-based cluster analysis presented below. Since clustering methods are typically influenced by the presence of outliers, data is transformed with the goal to reduce skewness and force equivalent variable ranges. The latter is also an important requirement for cluster analysis, as variables with

[15] Company valuations are the product of VC investment valuations — mostly reported following Invest Europe guidelines — and the stakes acquired via the investment. Patents are sourced from Orbis/PATSTAT.

[16] To minimise data loss, the mid- and post-investment years are converted into biannual periods. For the "baseline" period, any data point available between one year prior to investment and one year after investment is selected, with the most "appropriate" chosen first, *e.g.* *pre-* before *post-*. Follow-up periods contain any non-missing data point within the horizon biennium. For instance, for the 5 yr. growth horizon, the follow-up statistic is the last non-missing observation among the 4^{th} and 5^{th} post-investment measurements.

[17] The 2011 investment cohort could have possibly been included as, time-wise, its 5-year horizon threshold is in most cases achieved. However, only rarely can 2016 data can be retrieved for such companies. In fact, the use of this cohort would bias the analysis, for only a few defaulted companies would be observed. Thus, the 2011 cohort is also excluded from the analysis.

larger ranges tend to disproportionately influence the clustering process. To address these issues, this work employs the *neglog* transformation (Whittaker *et al.*, 2005).[18] The use of such transformation is convenient as it approximates well the log-transformation for positive values, while at the same time allowing for negative- and zero-valued observations of CAGRs (occurring when start-up size declines or stagnates over periods). Resulting values are further standardised as in Delmar *et al.* (2003).

6 Results

The cluster analysis methodology presented above produces a solution based on four "growth clusters". In fact, additional clusters were identified: a five- and seven-clusters solution are also observed, providing a local optimum for the model selection criterion. However, beyond the four-group solution, further clusters only add smaller groups of non-informative outliers. Thus, the four-group classification better suits the scope of this research. Appendix F briefly discusses the other solutions.

6.1 Analysis of clusters

Table 1 portrays the key descriptive statistics for each of the four identified growth profiles.

Table 1: Average (and median) CAGR by growth profile

5yr. CAGR in terms of:	Growth profile (5yr. horizon) under-performers	moderate performers	sale-based growth	patent-based growth
Total assets	-89.55%	11.27%	76.07%	37.38%
	(-100.00%)	(5.12%)	(52.13%)	(22.41%)
Employees nr.	-86.50%	24.74%	46.61%	23.40%
	(-100.00%)	(9.10%)	(44.28%)	(18.92%)
Turnover	-56.84%	51.74%	513.22%	145.21%
	(-100.00%)	(33.50%)	(449.51%)	(65.93%)
Company Valuation	-52.56%	-23.94%	5.45%	12.22%
	(-42.78%)	(-8.80%)	(1.07%)	(22.22%)
Patents nr.	8.35%	8.69%	12.44%	62.03%
	(0.00%)	(0.00%)	(7.63%)	(58.49%)
% of portfolio companies	12.40%	55.46%	12.14%	20.00%

Note: Medians reported in brackets. Statistics computed using *response propensity weights* (Little, 1986). N = 440.

Based on the findings of Table 1, the four clusters can be labelled as follows:

- *Under-performers*, representing almost 13% of the portfolio. In the first five post-investment years, these start-ups experience mostly negative growth rates, where the median rate is -100% (*i.e.* complete extinction). Company valuation reduces to about 3% of its original value, and while the average patenting growth rate is brought up to 8% by a few outliers, most under-performers in fact show no patenting activity at all.

[18] The neglog transformation is defined as:

$$\mathrm{nl}(x) = \begin{cases} -\log(-x+1) & x \le 0 \\ \log(x+1) & x > 0 \end{cases}$$

- *Moderate performers*, representing 55% of the portfolio. Five years after the investment date, these companies experience positive growth in all size measures: the median total assets increase is 25%, while employees rise is 45%. Turnover increases fourfold, as per the median CAGR. Despite significant growth, company valuation decreases by almost one fifth on median and down to 40% on average. Patenting growth rates are similar to the under-performing class, with a positive average but a null median CAGR.
- *Sale-based growers*, representing about 12% of the portfolio. This particular type of outperforming companies achieve an explosive 5-year growth driven by sales, accompanied by assets and employees increasing ninefold and sevenfold respectively. Interestingly, company valuation is predominantly stable, while patenting growth rates are positive and significant but orders of magnitude smaller than size variables.
- *Patent-based growers*, representing 20% of the portfolio. These companies show the highest patenting growth rates (median and average CAGR around 60%). Size growth is in-between moderate performers and sale-based growers, with assets and employees growing more than twofold. Patent-based growers experience the highest valuation growth rate, which also more than triples five years after investment.

Figure 5 uses a series of Box-and-whiskers plots to highlight major differences among the analysed growth profiles. For each growth pattern, Figure 5 provides ranges for the variables used in the analysis, emphasising the distinctive features of growth profiles (*e.g.* under-performers concentrating on the left side of the chart, turnover CAGR for sale-based growers rightmost to all other variables).

The remarkable growth rates experienced by outperforming start-ups cast a doubt on the fact that high-growth companies may be taking advantage of base effects to yield inflated 5-year CAGR. Thus, one potential distortion of such approach may be that while moderate performers will certainly grow less in relative terms, they may on the other hand grow more in absolute terms (*i.e.* generating more employment, assets and sales). Birch (1979) provides a convenient measure, since then named the *Birch index*, which helps countering the fallacy of relative growth analysis. In its simplest form, the index is obtained by multiplying relative growth with absolute growth.[19] Descriptive statistics on Birch indexes are portrayed in Table 2.

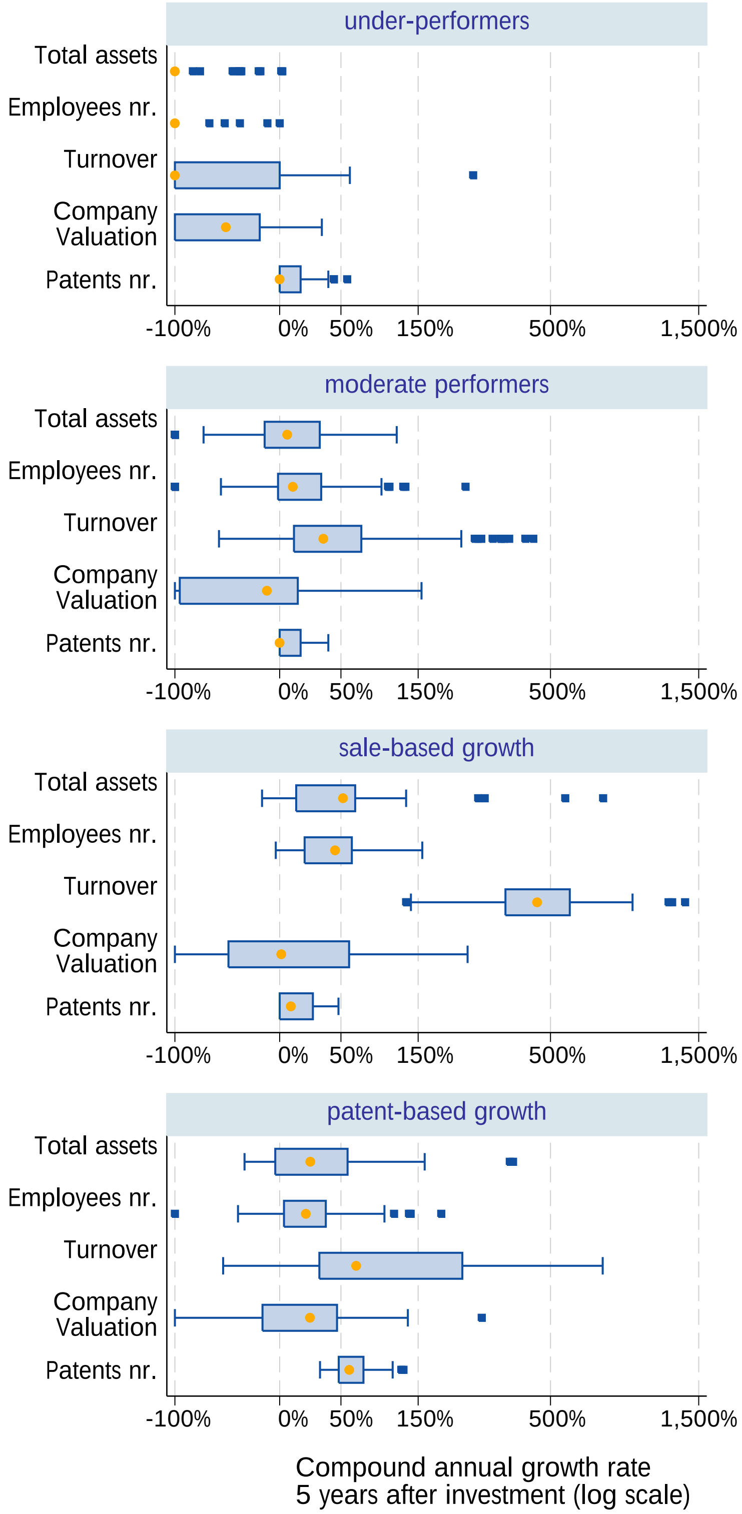
Table 2 offers interesting insights. On the one hand, it mostly validates the initial analysis on CAGRs. Indeed, the distinctive traits of both outperforming profiles prove to be superior to all other profiles also in terms of the median and average Birch index. However, it is also possible to note how some outliers cause the average Birch index of moderate performers to vastly overshadow median values. In conclusion, while the concern that relative growth rates portray a biased story of firm growth is well-founded, in reality it fails to drastically alter the results of this analysis.

6.2 Determinants of growth profiles

This section underlines the distinguishing features of the identified growth profiles. Following the canons of cluster analysis, a common approach to highlight differences among clusters is to assess

[19] In this work Birch indexes are computed by multiplying each 5-year CAGR with the (absolute) value at period end, expressed either in EUR million or units, depending on the indicator type.

Figure 5: Box-and-whisker plot of CAGRs by growth profile and clustering variable



Note: Orange dots indicate medians, boxes indicate interquartile ranges. Whiskers contain values distant at most 1.5 times the interquartile range (from the closest quartile). Small squares indicate outliers. All statistics computed using *response propensity weights* (Little, 1986). N = 440.

Table 2: Average (and median) Birch index by growth profile

5yr. Birch Index in terms of:	Growth profile (5yr. horizon) under-performers	moderate performers	sale-based growth	patent-based growth
Total assets	-0.14	23.14	15.94	8.94
	(0.00)	(0.19)	(1.56)	(1.24)
Employees nr.	-0.11	22.56	42.36	8.90
	(0.00)	(1.90)	(16.43)	(3.27)
Turnover	-0.02	4.20	129.45	4.18
	(0.00)	(0.84)	(20.82)	(1.13)
Company Valuation	0.31	10.72	100.30	444.14
	(0.00)	(0.00)	(0.18)	(1.79)
Patents nr.	0.53	0.87	1.38	11.28
	(0.00)	(0.00)	(0.59)	(6.66)
% of portfolio companies	12.40%	55.46%	12.14%	20.00%

Note: Medians reported in brackets. Statistics computed using *response propensity weights* (Little, 1986). N = 440.

whether typical features of start-ups — the so-called *passive variables*, where passive hints at the fact that they do not actively influence the clustering process — reveal significant discrepancies across different clusters. In other words, by comparing the distribution of passive variables against the general portfolio distribution, it is possible to single out potential determinants of growth profiles.[20]

To simplify the exercise, the analysis is limited to categorical determinants: these are sourced from the set of explanatory variables discussed in Coad (2007). Namely, geographic, sectoral, macroeconomic and age-related determinants. Numerical results are discussed in Appendix F, while Figure 6 provides a visual interpretation of these findings.

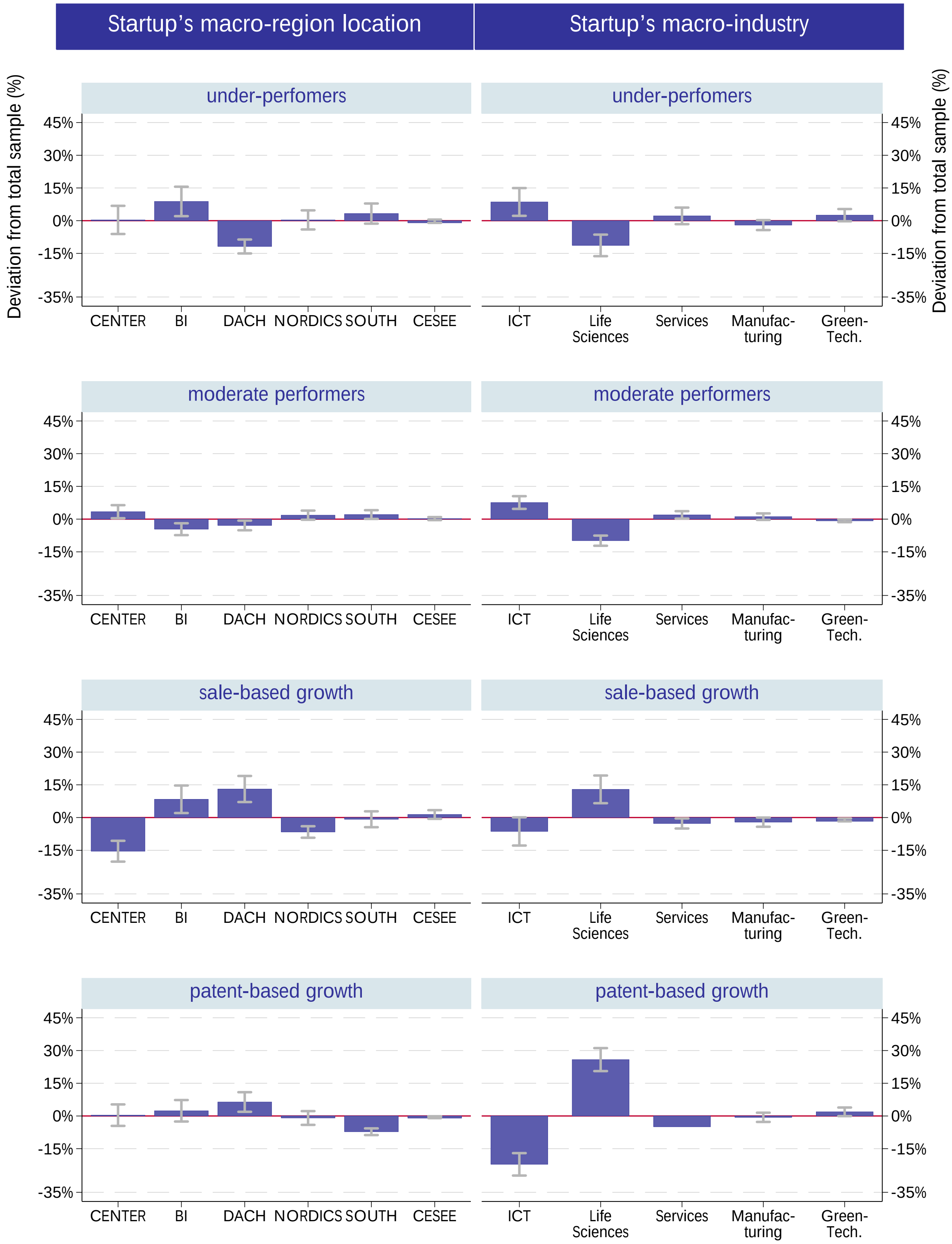
Figure 6 portrays, for all characteristics of each growth profile, their deviation from the overall portfolio proportion. For instance, a "surplus" of ICT companies in the under-performers profile indicates that for such profile companies tend to belong disproportionately more to this industry. Likewise, the "deficit" of life science start-ups signals that this sector produced less under-performers. All bars are complemented by 95% confidence intervals denoting the uncertainty level of each estimate.

Figure 6 offers numerous insights. Starting from regional areas, it portrays how growth profiles are evenly spread across regions, albeit with some significant differences. For instance, France- and Benelux-based startups find it harder to experience a sale-based explosive growth, which seems instead a feature of companies located in the British Isles. DACH start-ups contribute disproportionately less to the under-performer and more to the sale-based category. Companies located in the Nordic countries appear to be facing a lack of sale-based growers with respect to the other regions, while Southern European start-up face a lack of patent-based growers. Finally, no significant deviation can be noticed for CESEE companies.[21]

[20] This process is yet again impaired by the presence of missing data, whose effects are countered via the use of *response propensity weights* (Little, 1986). However, the original distribution is not always perfectly restored by the estimated weights. For this reason, the re-weighted sample is used to approximate the overall portfolio distribution.

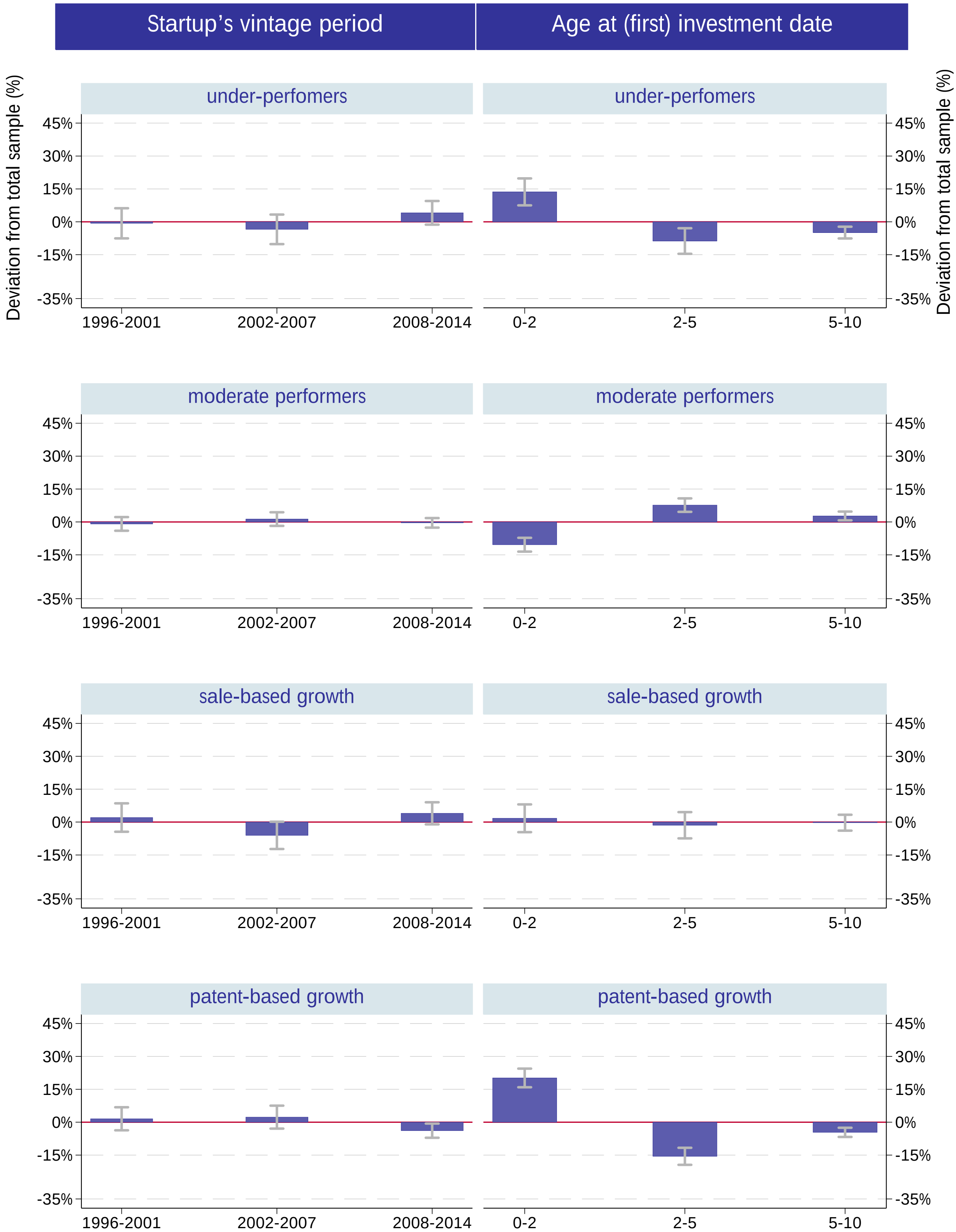
[21] However, given the relatively small proportion of such companies in the EIF portfolio, exacerbated by the fact that most of these have been invested in recent years omitted from the analysis (see Figure 1a), it may be too early to document the growth patterns occurring in this region.

Figure 6: Differences across growth profiles



Note: the figure above shows the distributional differences across subset of the portfolio, portrayed as deviations from the population proportion. All statistics computed using *response propensity weights* (Little, 1986). Re-weighting may occasionally fail to fully restore the original population distribution, so deviations are computed from overall sample data.

(Figure 6 continued)



Note: the figure above shows the distributional differences across subset of the portfolio, portrayed as deviations from the population proportion. All statistics computed using *response propensity weights* (Little, 1986). Re-weighting may occasionally fail to fully restore the original population distribution, so deviations are computed from overall sample data.

As per the industry determinants, ICT and Green-Tech companies appear disproportionately more likely to be under-performers than life science companies. While ICT companies offset this by showing a higher propensity to be moderate performers, Green-Tech startups show a surplus of patent-based growers.[22] ICT companies are also less likely to belong to the patent-based growers category: this may not be a surprising finding considering that ICT start-ups may have lower incentives to patent their innovations, compared to *e.g.* life science start-ups. These latter are indeed more likely to belong to both types of out-performers (and particularly patent-based growers) while also being less likely to be moderate- and under-performers. Service start-ups disproportionately have more moderate than explosive growth, while manufacturing companies do not seem to deviate substantially from the overall portfolio.

Concerning startup vintages, *i.e.* year of company first investments, Figure 6 evidences how these cohorts appear rather homogeneous. Some interesting findings that this analysis can only hint at concern recent investment cohorts: while they timidly show a surplus of under-performers, this phenomenon comes hand in hand with an additional share of sale-based growers, signalling that these companies may simply be riskier. Conversely, the 2002–2007 investment period benefits from a less risky, lower rate of under-performers.

The age distribution provides insights that are in line with the relevant literature. Younger companies, aged less than 2 years, tend to be disproportionately more under-performers. At the same time, these companies also show higher rates of patent-based growers. All in all, younger companies also appear riskier: their growth performance is considerably less predictable than "older" ventures, whichever the direction of such difference. On this point, it can be yet again remarked how older companies discount their relative shortage of under-performers with a surplus of moderate performers, as well as a relative lack of out-performing companies.

Among other potential determinants of growth profiles, the timing and size of VC investments may play a significant role. While a detailed exploration of this subject lies beyond the scope of this paper, results indicate that average investment levels tend to be smaller for under-performers and bigger for patent-based growers. Interestingly, moderate and sale-based performers experience investment levels that are not significantly different from the rest of the portfolio. Similar findings are observable for the time span of the investment: under-performers face significantly shorter holding periods, patent-based growers face significantly longer ones, while moderate and sale-based performers are, on average, in line with the overall portfolio.

6.3 Growth profiles as predictors of start-up success

A possible argument against the findings described thus far could challenge the view that 5-year CAGRs provide an accurate portrayal of the long-term growth of start-ups. To tackle this argument, a two-step strategy is presented: first, this section analyses the extent to which the observed growth profiles are stable over time. Following that, the analysis moves its focus to the exit performance of the observed clusters.

[22] Given the small number of green-tech startups, the same disclaimer given in footnote 21 applies.

The model-based cluster analysis introduced in section 5 allows to compute transitioning rates between growth profiles across subsequent time horizons. First, the 5-year horizon is compared to the growth in the first 3 years, to assess whether early signs of success and/or failure are observable among EIF-backed startups. Second, 5-year horizon profiles are compared against growth patterns 7 years after investment, providing some indications on the degree to which the 5-year horizon can predict longer-term trends. To simplify the discourse, it is convenient to define the 3rd, 5th and 7th post-investment year respectively as the *short*, *medium* and *long* term.

Table 3 includes the observed transition probabilities for companies whose short- and medium-term growth rate could be computed. If the company is confirmed inactive in its 5th post-investment year, it is kept in the sample and assigned the medium-term status *defaulted*. This allows to account for the occurrence of defaults and partly addresses the issue of survivorship bias. As it has been customary throughout this work, statistics are re-weighted to counter the bias brought by missing data.

Table 3: Growth profile transition (3 years *vs* 5 years after investment)

Growth profile (3yr. horizon)	Growth profile (5yr. horizon) under-perfomers	moderate performers	sale-based growth	patent-based growth	defaulted
under-perfomers	45.22%	31.05%	0.00%	0.00%	23.73%
moderate performers	1.80%	93.66%	0.71%	0.34%	3.49%
sale-based growth	0.00%	37.65%	46.94%	13.23%	2.18%
patent-based growth	4.84%	33.52%	3.33%	57.41%	0.90%

Notes: Statistics computed using *response propensity weights* (Little, 1986). N = 303.

Table 3 provides relevant insights. First, it is broadly noted how growth profiles tend to persist across the short- to medium-term: companies in each of the short-run categories were more likely to hold on to their pattern rather than transition to a new one in the medium term. However, there are some significant differences: while moderate performers represent the most stable classification, one third of sale-based and patent-based short-run outperformers are observed to shift downward to a moderate medium-term trend. Conversely and interestingly, under-performers show a similar rate of *upward* shifting in the medium-term. However, according to Table 3 two thirds of under-performers will still be facing difficulties in the medium-run, as one out of four under-performers is likely to default by the 5th year after investment. Looking at the longer-run growth, the perspective offered by Table 4 does not change manifestly, except for a few remarkable findings.

Table 4: Growth profile transition (5 years *vs* 7 years after investment)

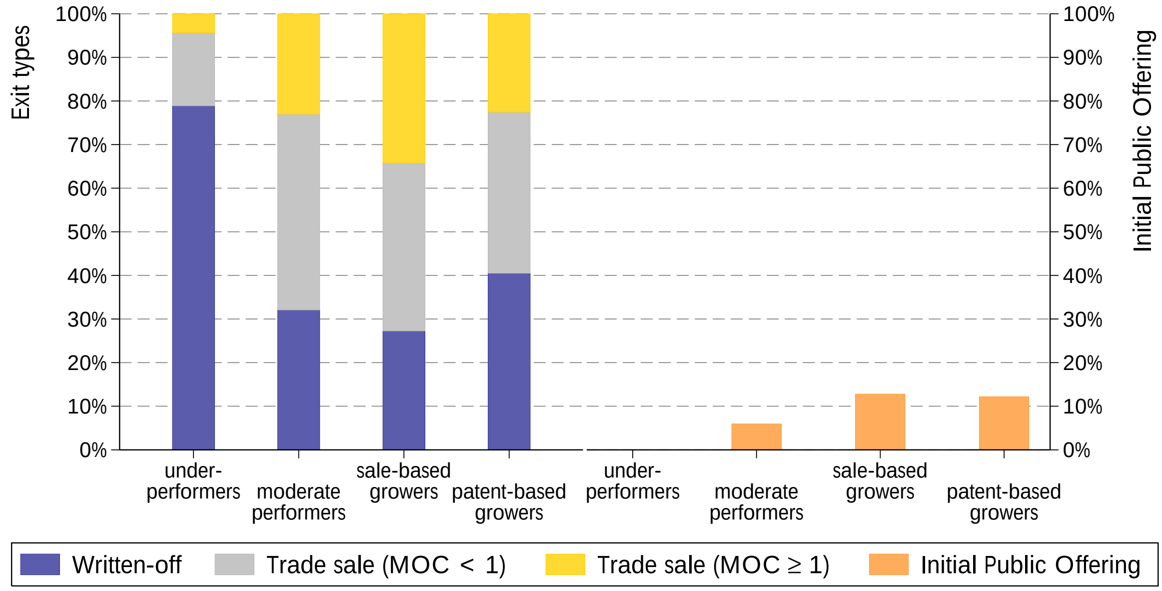
Growth profile (5yr. horizon)	Growth profile (7yr. horizon) under-perfomers	moderate performers	sale-based growth	patent-based growth	defaulted
under-perfomers	19.06%	70.14%	0.00%	0.00%	10.80%
moderate performers	0.56%	76.65%	10.08%	0.39%	12.33%
sale-based growth	0.00%	45.85%	50.40%	2.42%	1.33%
patent-based growth	0.00%	26.35%	2.75%	60.69%	10.21%

Notes: Statistics computed using *response propensity weights* (Little, 1986). N = 318.

Table 4 shows how start-ups with under-performing growth trends that are still kept in the portfolio face high chances of shifting to a moderate-performer classification. On the one hand, this finding points at the fact that the inherent long-term perspective of VC investments may eventually pay off for surviving companies.[23] However, under-performers are also never observed to shift from an under-performing pattern to an out-performing one, a feature that restrains the former pay-off. With regards to moderate medium-term performers, transition rates show that only 10% of these are expected to shift to out-performing statuses, while about the same rate is likely to face default.

Overall, transition rates show a certain degree of path-dependency among profiles of growth. Although it is certainly possible for VC-backed companies to transition to higher growth speeds, in fact most start-ups remain faithful to their initial growth pattern. Moreover, Table 3 and 4 evidence a tendency for start-ups to converge to a "moderate" pattern of growth. Although far from insuccessful, Table 1 has shown how moderate growth is associated with a decrease in the valuation of the company. The extent to which this finding may be of concern to VC investors is questionable, as lowering valuations do not necessarily imply losses. To address this question, Figure 7 portrays the exit performance and the initial public offering (IPO) rate experienced by each start-up profile.

Figure 7: Exit performance and IPOs by 5-year growth profile



Note: Exit performances only refer to the portion of fully exited EIF-backed startups. To appear in the statistics, start-ups must have fully exited all EIF-backed VC funds. Initial public offerings, on the other hand, can occur also while companies are still being actively invested. Therefore, the two series do not overlap. Statistics computed using *response propensity weights* (Little, 1986). N = 440.

Figure 7 portrays the major exit attainments of each growth profiles, also providing evidence on the predictive ability of medium-term growth against exit performance.[24] First, it can generally be

[23] This pay-off is not automatically transferred to the investor, as further discussed.

[24] As Figure 7 is the result of aggregated performances over the last 20 years, an important disclaimer is due here: exit performances are inherently linked to the macroeconomic environment. The performance portrayed above may not be reflective of more recent trends. Forthcoming issues in the series will focus on EIF-backed exits to provide a comprehensive account of investment performances over time.

observed that higher-performing companies do, on average, provide better exit opportunities. On the other hand, medium-term under-performers face a 80% chance of being written-off, with only one out of 20 invested companies generating positive investment returns. Second, sale-based and patent-based growers generate better investment opportunities, but while exits of sale-based companies outperform all other profiles, patent-based growers appear not significantly different from moderate performers. In fact, throughout the observed period patent-based start-ups have faced higher write-off rates than moderate performers. However, this is primarily due to the aftermath of the dot-com crash, when both sale-based and patent-based growers suffered higher rates of write-off and below-return trade sales than moderate performers. In recent periods, high-growth start-ups significantly outperformed moderate growers in terms of exit class. Third, IPO rates tend to comply with expectations, that is, higher-growth companies benefit from a higher probability of going public.

Overall, Figure 7 also evidences the low predictive ability of medium-term growth profiles. While negative size growth rates are certainly a good predictor of a company's write-off probability, investments in moderate- and high-growth start-ups still show high chances of not being profitable. The task to identify more appropriate determinants of profitable exits is left to future works in this area.

6.4 Further insights on the geography of growth profiles

This section concludes the analysis of growth profiles by further discussing the geographical aspects of growth performance. In section 6.2, regions were observed to have fairly homogeneous results in terms of under- and moderately-performing start-ups. However, significant differences were noted with respect to the lack or surplus of patent-based and sale-based high-growth companies. To shed further light on such findings, the methodological approach employed thus far cannot be pursued, as the high degree of missing information makes it virtually impossible to generate a sample that is faithful to the geographic properties of the EIF VC portfolio.

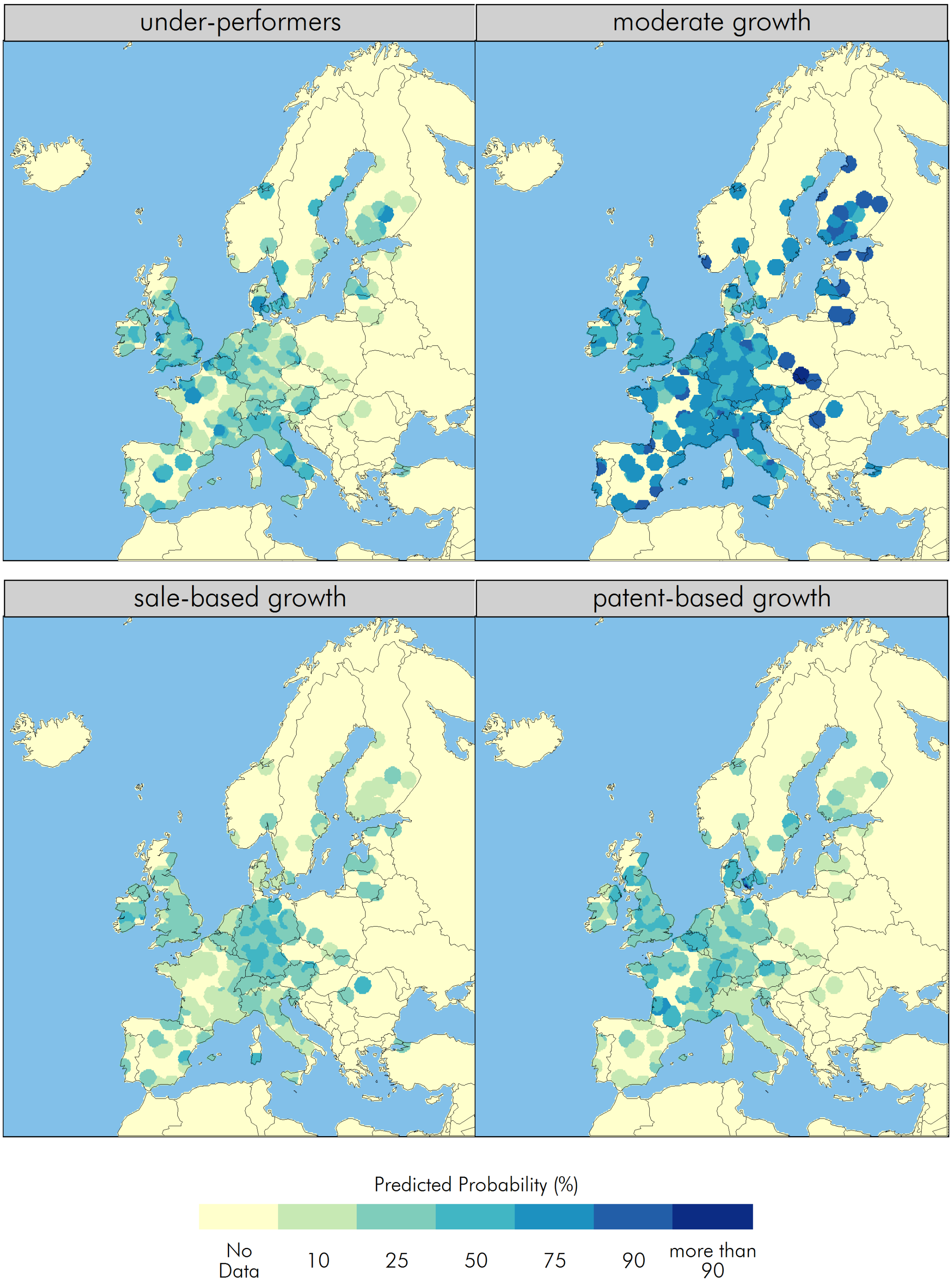
Against this background, the following analysis uses a different approach: exploiting the virtue of the parametric clustering approach introduced in section 5, probabilities of being affiliated to each of the four identified growth profiles are regressed on several explanatory variables that are observable for the entirety of the EIF VC portfolio. Issues related to selection on unobserved variables discussed in section 3 are addressed through appropriate model design, documented in Appendix H. This allows to estimate, for each portfolio company, the probability of its affiliation to any identified growth profile. Aggregated values at the city-level offer important insights on the geographic spread of growth patterns.

Granted, this is an exercise in extrapolation: results can only be considered as indicative of the true underlying phenomena. Nevertheless, there is an inherent value in assessing whether growth profiles tend to be concentrated in key regional areas. Results of this analysis are portrayed in Figure 8, which portrays the predicted probabilities for each growth profiles across regions of Europe.[25]

A number of interesting findings arise from Figure 8. Concerning the distribution of under-performers, the findings of section 6.2 are broadly confirmed: no entire region with a higher concentration of

[25] The reader is invited to compare this figure with the geographic dispersion of EIF-backed investments, shown in Figure 6 in Kraemer-Eis *et al.* (2016).

Figure 8: Geographic distribution of growth profiles probabilities



Note: probability maps for each profile are created from city-level averages of cluster affiliation probabilities. A detailed description of the approach is provided in Appendix H. The geographic distribution is estimated via quartic kernel function and a fixed bandwidth or approx. 17,500 km^2, *i.e.* the area of a circle with diameter 150km. All maps were created using the software in Pisati (2007).

such profile can be observed. However, there appears to be a higher probability of under-performers around key European hubs: further research may be necessary to confirm this finding, but one potential explanation relates to the fact that VC hubs tend to host companies with a lower age at investment date, a feature that has been shown in section 6.2 to be related to the higher incidence of riskier start-ups. As per the geographical features of moderate growers, no specific trend emerges from Figure 8. Aside from a number of sparse areas with high concentration rates, typically due to low sample sizes, moderately growing start-ups are roughly evenly spread across all geographies.

6.4.1 Sale-based *vs* patent-based growth

The most insightful results of Figure 8 perhaps pertains to the geographical spread of sale-based and patent-based growers. Indeed, the portrayed maps highlight how these high-growth profiles tend to concentrate in different geographical areas, with a more limited overlap than what can be observed for other profiles. To confirm this finding, Figure 9 portrays an indicator that seeks to single out areas with a higher propensity towards one of the two high-growth profiles.

The indicator builds on the Birch index discussed in section 6.1, and weights the difference between sale-based and patent-based grower predicted companies by the relative proportion of out-performers observed in a given area. This approach corrects for the case in which very few out-performing investments can shape the growth bias of lager areas. Additional details are discussed in Appendix H.

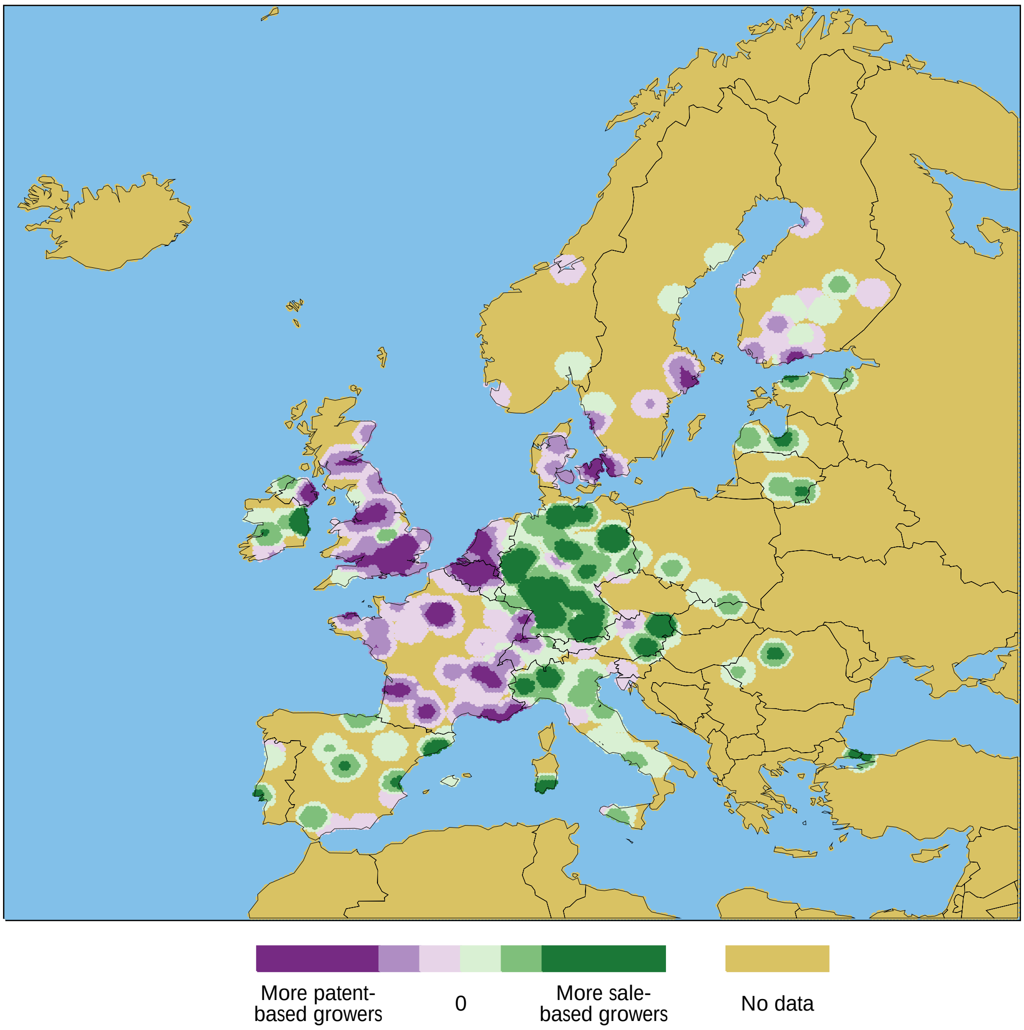
Figure 9 provides interesting results. Across Europe, the EIF VC activity highlights the bias of some hubs towards sale-driven growth (*e.g.* in Berlin, Munich, Milan, Dublin). Conversely, other hubs seem to be more specialised in patent-driven growth (*e.g.* Paris, London, Amsterdam). While the broad picture appears to evidence significant specialisation at the country level, it is possible upon close inspection to identify in most countries regions specialised in patent-driven growth, as well as others yielding more sale-driven high-growers. In conclusion, it is important to recall that the main difference between sale-driven and patent-driven growers lies not only in their sectoral affiliation: while it is true that patent-based growers disproportionately operate in the life science segment, 40% of these are actually ICT companies. Similarly, 35% of sale-based growers operate in the life science segment. The findings portrayed in this section are thus concerned with the broader concept of growth pattern, and Figure 9 perhaps highlights the role of macro-economic factors (regional and/or national) to act as *enablers* of different growth trajectories. On this point, further research in this area is certainly needed prior to reaching conclusive evidence.

7 Conclusions

Start-up growth is often treated as a stylised fact. However, as discussed in section 2, such perspective disregards an extensive literature characterised by divergent theories and empirical findings. Against this background, this work contributes to the literature on start-up growth by analysing a specific subset of European technology-based start-ups, supported by EIF-backed VC investments.

Employing a wide range of descriptive statistics, the analysis portrayed in section 4 documents the significant growth of EIF-backed start-ups in the aftermath of a VC investment, both on average and

Figure 9: Geographic bias of high-growth profiles



Note: the bias index is obtained by multiplying the surplus of sale- or patent-driven growers by the relative proportion of out-performers in a given area. A formal description of the index is provided in Appendix H. From left to right, the distribution classes based on the 10th, 25th, 50th, 75th and 90th percentile respectively. All maps created using the software in Pisati (2007).

median terms. In particular, it is observed that, on average, EIF-backed start-ups experience at least a twofold increase in their number of employees and total assets within four years after investment date. Moreover, it is noted that seven years since investment date, most profitability indicators and ratios hint at positive trajectories of such EIF-backed companies. For instance, Figure 4 in section 4.2 shows how in a 7-year horizon the proportion of firms with positive ROA raises from 10% to 35%. The analysis of growth trends addresses missing data concerns by using a robust re-weighting strategy based on Little (1986), Van de Ven and Van Praag (1981) and Haziza and Beaumont (2007).

However, the remarkable findings evidenced in section 4 are to be weighted against the pervasiveness of survivorship bias, which filters out defaulting companies and accentuates the performance of successful start-ups. Future works under this Working Paper series will address this important phenomenon and enhance the knowledge set on start-up medium- to long-term development. Moreover, the insights evidenced by descriptive statistics are perturbed by the presence of extreme outliers and, in general, the high heterogeneity of growth trends. Outliers are a defining feature of the venture capital industry, whose effects cannot be dismissed. Therefore, the mere use of descriptive statistics renders difficult the identification of one "typical" growth trend of start-ups, as various appear to exist.

For this reason, section 6 discusses the results of a *latent class cluster analysis* (Skrondal and Rabe-Hesketh, 2004), a form of cluster analysis based on a parametric model. By looking at 5-year start-up growth rates, four main profiles of growth are identified: *a) under-performers*, representing almost 13% of the portfolio, *b) moderate performers*, constituting 55% all investees, and two type of out-performers. These are *c) sale-based growers* and *d) patent-based growers*, representing 12% and 20% of the portfolio respectively.

Analysing the distribution of growth profiles, section 6.2 identifies key regional, sectoral and age-related features associated with an increased presence of each growth profile. For instance, European life science start-ups have a higher propensity than ICT companies to become both sale-driven and patent-driven out-performers.[26]

Growth profiles tend to be persistent over time: in most cases, it is more likely that start-ups hold on to their profile than transition to a new one. If convergence towards a certain state is observed, then it typically leads to more moderate growth. Moreover, section 6.3 shows how medium-term growth rates have a low predictive ability towards investments returns: while out-performing companies are on average more remunerative at exit, all profiles still face a substantial risk of investment write-off, which is at best in the range of 30%.

The paper concludes by further exploring the geographical dimension of growth profiles: by predicting the growth pattern of each EIF-backed start-up, section 6.4 evidences regions of Europe where certain growth profiles are dominant, in particular with regards to out-performing companies. The analysis documents the existence of several different geographical clusters in which one of these two growth modes dominates the other. As growth profiles tend to cluster around key areas (*e.g.* hubs) it is speculated that regional and/or national macro-economic factors — acting as *enablers* of different growth trajectories — play a significant role in the growth mode of start-ups. Further research could build on the growth profiles identified in this work and look into the determinants of sale- *versus* patent-led growth, in particular with respect to their business models. Overall, the findings highlight the potential for EIF-backed VC start-ups to significantly contribute to the economic development and job creation across several regions of Europe.

Forthcoming issues in the Working Paper series will focus on aspects left mostly untouched. Additional work in the area of investment returns, innovation, survivorship and job creation of companies will be necessary in order to draw a comprehensive representation of EIF-backed start-ups. In particular, further efforts in the area of employment growth may be needed, as one limit of *Orbis* data concerns the provider's inability to track off-payroll employees, whose impact on total start-up workforce may prove significant. The series will hence conclude with an assessment of EIF's impact on start-ups it has supported.

[26] Although, in absolute terms, the higher proportion of ICT companies in the portfolio causes sale-driven ICT growers to outnumber those in life science.

References

Almus, M. and Nerlinger, E.A. (1999). Growth of new technology-based firms: Which factors matter? *Small Business Economics*, vol. 13, no. 2, pp. 141–154.

Berk, R.A. (2008). *Statistical Learning from a Regression Perspective*. Springer-Verlag New York, 1 ed.

Bertoni, F., Colombo, M.G. and Grilli, L. (2011). Venture capital financing and the growth of high-tech start-ups: Disentangling treatment from selection effects. *Research Policy*, vol. 40, no. 7, pp. 1028 – 1043. Issue dedicated to Chris Freeman, Founding Editor.

Bethlehem, J., Cobben, F. and Schouten, B. (2011). *Handbook of Nonresponse in Household Surveys*. Wiley.

Birch, D.L. (1979). *The job generation process*. MIT Program on Neighborhood and Regional Change, Cambridge, MA: Massachusetts Institute of Technology.

Bronzini, R., Caramellino, G. and Magri, S. (2015). Venture Capitalists at Work: What Are the Effects on the Firms They Finance? Working paper, Bank of Italy, Rome.

Calvino, F., Criscuolo, C. and Menon, C. (2015). Cross-country evidence on start-up dynamics. OECD Science, Technology and Industry Working Papers 2015/06, OECD Publishing, Paris.

Coad, A. (2007). Firm growth : a survey. Documents de travail du Centre d'Economie de la Sorbonne 2007.24 - ISSN : 1955-611X.

——— (2009). *The Growth of Firms A Survey of Theories and Empirical Evidence*. Edward Elgar Publishing, Inc., Cheltenham, UK.

David, M., Little, R., Samuhel, M. and Triest, R. (1983). Imputation models based on the propensity to respond. *Proceedings of the Business and Economics Section*, pp. 168–173.

Davila, A., Foster, G. and Gupta, M. (2003). Venture capital financing and the growth of startup firms. *Journal of Business Venturing*, vol. 18, no. 6.

Delmar, F., Davidsson, P. and Gartner, W.B. (2003). Arriving at the high-growth firm. *Journal of Business Venturing*, vol. 18, no. 2, pp. 189 – 216.

Dubin, J.A. and Rivers, D. (1989). Selection Bias in Linear Regression, Logit and Probit Models. *Sociological Methods & Research*, vol. 18, no. 2-3, pp. 360–390.

Everitt, B. (2011). *Cluster analysis*. John Wiley & Sons, 5 ed.

Garnsey, E. and Heffernan, P. (2005). Growth setbacks in new firms. *Futures*, vol. 37, no. 7, pp. 675 – 697. Complexity and the limits of knowledge.

Gilbert, B.A., McDougall, P.P. and Audretsch, D.B. (2006). New venture growth: A review and extension. *Journal of Management*, vol. 32, no. 6, pp. 926–950.

Haziza, D. and Beaumont, J.F. (2007). On the construction of imputation classes in surveys. *International Statistical Review*, vol. 75, no. 1, pp. 25–43.

Heckman, J.J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, vol. 47, no. 1, pp. 153–161.

Kazanjian, R.K. and Drazin, R. (1990). A stage-contingent model of design and growth for technology based new ventures. *Journal of Business Venturing*, vol. 5, no. 3, pp. 137–150.

Kraemer-Eis, H., Signore, S. and Prencipe, D. (2016). The European venture capital landscape: an EIF perspective. Volume I: the impact of EIF on the VC ecosystem. EIF Working Paper 2016/34, EIF Research & Market Analysis.

Little, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review / Revue Internationale de Statistique*, vol. 54, no. 2, pp. 139–157.

Lünnemann, P. and Mathä, T.Y. (2001). Monetary transmission: empirical evidence from Luxembourg firm level data. Working Paper Series 0111, European Central Bank.

McKelvie, A. and Wiklund, J. (2010). Advancing firm growth research: A focus on growth mode instead of growth rate. *Entrepreneurship Theory and Practice*, vol. 34, no. 2, pp. 261–288.

Nicoletti, C. and Peracchi, F. (2005). Survey response and survey characteristics: microlevel evidence from the European Community Household Panel. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 168, no. 4, pp. 763–781.

Papke, L.E. and Wooldridge, J.M. (1996). Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics*, vol. 11, no. 6, pp. 619–632.

Pisati, M. (2007). SPMAP: Stata module to visualize spatial data. Stata Help Files. URL https://ideas.repec.org/c/boc/bocode/s456812.html.

Puri, M. and Zarutskie, R. (2012). On the Life Cycle Dynamics of Venture-Capital- and Non-Venture-Capital-Financed Firms. *The Journal of Finance*, vol. 67, no. 6, pp. 2247–2293.

Robb, A.M. and Robinson, D.T. (2014). The capital structure decisions of new firms. *Review of Financial Studies*, vol. 27.

Rosenbaum, P.R. and Rubin, D.B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, vol. 70, no. 1, pp. 41–55.

Rubin, D.B. (1976). Inference and Missing Data (with discussion). *Biometrika*, vol. 63, no. 3, pp. 581–592.

Schonlau, M. (2002). The clustergram: A graph for visualizing hierarchical and nonhierarchical cluster analyses. *Stata Journal*, vol. 2, no. 4, pp. 391–402.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Chapman and Hall/CRC, 1 ed.

Van de Ven, W.P. and Van Praag, B.M. (1981). The demand for deductibles in private health insurance. *Journal of Econometrics*, vol. 17, no. 2, pp. 229–252.

Welch, C.A., Petersen, I., Bartlett, J.W., White, I.R., Marston, L., Morris, R.W., Nazareth, I., Walters, K. and Carpenter, J. (2014). Evaluation of two-fold fully conditional specification multiple imputation for longitudinal electronic health record data. *Statistics in Medicine*, vol. 33, no. 21, pp. 3725–3737.

Whittaker, J., Whitehead, C. and Somers, M. (2005). The neglog transformation and quantile regression for the analysis of a large credit scoring database. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 54, no. 5, pp. 863–878.

Woodruff, R.S. (1952). Confidence Intervals for Medians and Other Position Measures. *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 635–646.

Appendices

A NACE classification and correspondence table

The EIF sectoral nomenclature is based on the 2007 sectoral classification of Invest Europe.[27] A total of 16 sectors results from the aggregation of the 24 more granular industries outlined by Invest Europe. Six "macro-sectors" are further derived by aggregating subsets of EIF sectors. The sectoral classification of Invest Europe also offers a concordance table with the NACE Rev. 2 statistical classification of economic activities,[28] allowing to link the EIF sectoral nomenclature with the NACE general classification. Table A1 shows the final correspondence table.

NACE sectoral divisions are used to identify the appropriate *producer price indices (PPIs)*, which are further employed to convert monetary values to constant Euros (base year 2005). Four main data sources are used to retrieve PPIs. These observe a "pecking order" organised as follows: Eurostat, OECD, World Bank (WB) and National Statistical Offices (NSOs). For EU28-based companies (*i.e.* about 91% of the EIF VC portfolio as of end-2014), Eurostat PPIs by 10 industry branches provide 75%, OECD 15% and NSOs 3.5% of the necessary data points. WB data (referring to overall economy prices) covers additional 3.5% observations, while the remainder is imputed via predictive mean matching. For *non* EU28-based companies, PPI data is mostly sourced from WB and OECD.

Table A1: Correspondence table between EIF sectors of activity and NACE Rev. 2 classification

Macro-sector classification	EIF sectors of activity	NACE rev.2 division	NACE 10-branches class
ICT	Communications	J	Information and communication
ICT	Computer Related	J	Information and communication
ICT	Industrial Automation	B-E	Industry (except construction)
ICT	Other Electronic Related	J	Information and communication
Life Sciences	Biotechnology	M-N	Professional, scientific and technical activities; administrative and support service activities
Life Sciences	Medical/health Related	O-Q	Public administration, defence, education, human health and social work activities
Manufacturing	Chemicals and Materials	B-E	Industry (except construction)
Manufacturing	Construction	F	Construction
Manufacturing	Industrial Products and Services	B-E	Industry (except construction)
Manufacturing	Other Manufacturing	C	Manufacturing
Services	Consumer Related	G-I	Wholesale and retail trade, transport, accomodation and food service activities
Services	Financial Services	K	Financial and insurance activities
Services	Other Services	R-U	Arts, entertainment and recreation; other service activities; activities of household and extra-territorial organizations and bodies
Services	Transportation	G-I	Wholesale and retail trade, transport, accomodation and food service activities
Green Technologies	Agriculture	A	Agriculture, forestry and fishing
Green Technologies	Energy	B-E	Industry (except construction)

[27] Available at http://www.investeurope.eu/media/12926/sectoral_classification.pdf

[28] NACE stands for *Nomenclature générale des Activités économiques dans les Communautés Européennes* (General Industrial Classification of Economic Activities within the European Communities).

B Re-weighting methods

The re-weighting approach used in this study borrows from nonresponse adjustment methods in the context of surveys. Nonresponse adjustment typically assigns weights to survey respondents on the basis of auxiliary information. With respect to the latter, assigned weights seek to make respondents representative of the population. This method assumes that nonresponse towards a target variable is *missing-at-random* (MAR, see Rubin, 1976). According to Bethlehem *et al.* (2011), the bias-reduction ability of this strategy depends on the *relevance* of auxiliary information, with respect to both the response behaviour and the target variable. One popularised strategy to nonresponse adjustment entails the use of *response propensity*.

Response propensity (David *et al.*, 1983) draws on the propensity score theory of Rosenbaum and Rubin (1983). Following Little (1986), suppose that y is a target variable and r is the response mechanism. For $\bar{y}_r$, the respondent mean for y, to be unbiased with respect to the population mean it is required that $y \perp r$, *i.e.* that y and r are independent. In this case, r is said to generate *missing completely at random* (MCAR) data. With MCAR data, the response mechanism is independent of y. In other words, consider $\mathbf{z}$ as the set of unknown parameters that affects the response probability: with MCAR the probability of individual i responding, conditional on y_i is

$$\mathrm{pr}\left(r_i = 1 | y_i, \mathbf{z}_i\right) = \mathrm{pr}\left(r_i = 1 | \mathbf{z}_i\right) \tag{1}$$

i.e. the mechanism at which data is retrieved is equivalent to simple random sampling (Berk, 2008). This assumption is not verifiable in the case of EIF-backed start-ups, as a number of determinants that disproportionately affect the response propensity can be observed. However, suppose that $\mathbf{a}$, a set of auxiliary variables, can completely remove the nonresponse bias

$$y \perp r | \mathbf{a} \tag{2}$$

define the *response propensity* as $p(\mathbf{a}) = \mathrm{pr}\left(r = 1 | \mathbf{a}\right)$. If $p(\mathbf{a}) > 0$ for all observed values of $\mathbf{a}$, then Rosenbaum and Rubin (1983) show that (2) implies

$$y \perp r | p(\mathbf{a}) \tag{3}$$

i.e. nonresponse bias can be completely removed via re-weighting on $\mathbf{a}$. Little (1986) proposes a two-step strategy to remove nonresponse bias with response propensity:

1. Estimate the response propensity $p(\mathbf{a})$ by regressing r on $\mathbf{a}$, obtaining $\hat{p}(\mathbf{a})$.

2. Form adjustment cells based on the inverse of $\hat{p}_D$, a *discretised* version of $\hat{p}(\mathbf{a})$.

Little (1986) describes this method as *response propensity stratification*, arguing that step 2 is necessary to avoid very small values of $\hat{p}(\mathbf{a})$ to inflate the variance of re-weighted estimates.

The dichotomous nature of response behaviour suggests the use of probit or logit models for the estimation of the response propensity $p(\mathbf{a})$. Bethlehem *et al.* (2011) note that while both approaches are mostly equivalent, it is however important to consider the nature of the response process when estimating $\mathbf{p}(\mathbf{a})$. In particular, the response behaviour can sometimes result from an elaborate, multi-step response process. With regards to *Orbis* data, two main causes of missing data can be identified: *a)* the company is not *matched*, i.e. its financial records cannot be found, and *b)* the company is not *usable*, i.e. its available data cannot be used to meaningfully derive growth trends.

This work introduces the concept of *usable* company, defined as follows: to be usable, companies must offer two or more data points: one in the *baseline* period and at least one in the *follow-up* period, to assess growth. The periods cannot overlap. The *baseline* is defined as the period occurring from one year before the first investment date, to one year after such date. The *follow-up* period starts from the first year after investment and adds up as long as the company is alive and actively invested.

The absence of *matching* for Orbis companies causes a selection bias for which company financials cannot be observed, hence *usable*. The selection bias is endogenous, for the error term in the selection (*matching*) equation is correlated with the error in the *usable* equation of interest. In this case, the probit model with sample selection (Van de Ven and Van Praag, 1981) can provide an appropriate parametrisation of the analysed response process. To be identifiable, such model requires an exclusion restriction, *i.e.* the existence of one or more variables affecting the selection process, but not the outcome of the equation of interest. Following Nicoletti and Peracchi (2005), this paper satisfies the exclusion restriction by using characteristics of the data provider and details on the collection of accounts of Orbis companies. These are sourced from Bureau Van Dijk's Orbis user guide and are provided on a country-by-country basis.

Additional auxiliary variables, non-missing for all portfolio companies, are used in the selection equation to control for various fixed-effect characteristics. The estimated coefficients for auxiliary variables are reported in Table H2 for some of the used variables.[29] These are used to compute the company-specific response propensity for each analysed variable. Despite different missing rates, the *response mechanism* seems to affect financial indicators in a rather homogeneous way, as hinted by the coefficients in Table H2. Two major arguments are provided for the choice to apply weights to companies as opposed to company-year pairs. From a practical perspective, company-years weights are observed to inflate the variance of the resulting weights. From a theoretical standpoint, company-level weights based on $\mathbf{a}$ are sufficient to offset the two main sources of sample attrition: company default (permanent) and non-collection (temporary). The first is a direct consequence of (3), while non-collection (*i.e.* "gaps" in the series for particular years) is assumed completely random.

The estimated response propensity $\hat{\mathbf{p}}(\mathbf{a})$ is now estimated. To proceed to its stratification, the researcher must design grouping classes. Two non-trivial questions thus need answered: *how* to select grouping classes and *how many* of these to impose. Haziza and Beaumont (2007) offer valuable anchor points in their Monte Carlo-based study. The authors find that grouping based on a *k-means* clustering algorithm consistently scores better than alternatives such as the creation of evenly spread classes. Moreover, the authors observe that between 10 and 25 classes typically provide an effective bias reduction, even in the case of a misspecified response estimation model.

[29] These exemplify over 100 models estimated via the same approach in various parts of this work.

Table B1: Response propensity estimation. Dependent variables on column names.

Variable	Empl.	Turnover	Total assets	ROA	Leverage ratio
usable equation					
Investment features:					
Investment amount at first date	0.000	0.000*	-0.000	0.000	-0.000
Investment year	-0.063***	-0.134***	-0.113***	-0.137***	-0.114***
Nr. of investment rounds	0.077***	0.061**	0.072***	0.050*	0.075***
Nr. of investor funds	-0.206**	-0.257***	-0.240***	-0.245***	-0.244***
Size of VC portfolio	0.001	0.001	0.001	0.002	0.001
Start-up experienced IPO†	0.348**	0.815***	0.477***	0.646***	0.501***
Start-up has high growth potential†	0.320***	0.253***	0.398***	0.289***	0.406***
VC firm located in hub†	0.259**	0.226**	0.122	0.206*	0.149
BvD collection:					
Time for records to appear in BvD	-0.149**	-0.067	-0.112*	-0.030	-0.090
Company matched via BvD algorithm†	0.127*	0.225***	0.382***	0.280***	0.357***
Economic environment:					
Country PPI	0.392	1.165**	-0.763	1.634***	-0.696
Standardised Financial records	Yes	Yes	Yes	Yes	Yes
Start-up investment period	Yes	Yes	Yes	Yes	Yes
Start-up macro-region	Yes	Yes	Yes	Yes	Yes
Start-up stage at investment	Yes	Yes	Yes	Yes	Yes
matched equation					
Investor features:					
Size of VC fund	-0.000***	-0.000**	-0.000**	-0.000**	-0.000**
Investment features:					
EIF share in supported investment	-1.485***	-1.546***	-1.569***	-1.545***	-1.554***
Investment amount at first date	0.000*	0.000*	0.000*	0.000*	0.000*
Investment year	0.089***	0.083***	0.084***	0.085***	0.083***
Nr. of investor funds	0.233	0.190	0.171	0.178	0.152
Nr. of quarters in the EIF portfolio	0.028***	0.026***	0.028***	0.027***	0.028***
Size of VC portfolio	-0.014***	-0.013***	-0.012***	-0.013***	-0.012***
Start-up has high innovation potential†	-0.113	-0.076	-0.005	-0.041	-0.009
Investee features:					
Age at investment	0.464***	0.482***	0.487***	0.467***	0.484***
Age squared at investment	-0.044***	-0.047***	-0.048***	-0.045***	-0.047***
BvD collection:					
Nr. of BvD data providers	0.057	0.108	0.128	0.116	0.126
Company matched via BvD algorithm†	2.141***	2.134***	2.099***	2.173***	2.100***
Companies comply with filing rules	Yes	Yes	Yes	Yes	Yes
Focus of VC fund	Yes	Yes	Yes	Yes	Yes
Institution storing financial records	Yes	Yes	Yes	Yes	Yes
Macro-region of VC fund	Yes	Yes	Yes	Yes	Yes
Start-up exit type	Yes	Yes	Yes	Yes	Yes
Start-up macro-industry	Yes	Yes	Yes	Yes	Yes
N	2951	2951	2951	2951	2951
Nr of matched	2679	2679	2679	2679	2679
Nr of usable	938	949	1375	960	1346

* $p<0.05$, ** $p<0.01$, *** $p<0.001$; † dichotomic variable; coefficients for constant term omitted.

To select the appropriate number of grouping classes, this work follows two additional heuristic approaches: *a)* bias reduction against observable auxiliary variables, and *b)* penalisation of groupings that generate extreme weights. This approach leads to the stratification over 11 classes for the descriptive statistics of section 4, and over 12 classes for the growth rates of section 5.

C List of financial indicators

Financial Indicator	Indicator dimension	Unit	Description	Formula
Nr. Of employees	Size	Unit	Total number of employees included in the start-up's payroll	
Turnover	Size	EUR million	Total operating revenues (incl. net sales, other operating revenues and stock variations); values do not include VAT	
Total assets	Size	EUR million	Total value of assets owned by the start-up	
Profit/Loss before taxes	Size/profit	EUR million	Results from the aggregation of all operating and financial revenues, minus all operating and financial expenses	
Return on assets	Profitability	Percent	An indicator of how profitable a company is relative to its total assets	$\mathrm{ROA} = \frac{\text{Profit before tax}}{\text{Total assets}}$
Pre-tax profit margin	Profitability	Percent	An indicator of the amount of revenue each EURO of turnover is generating	$\mathrm{PF} = \frac{\text{Profit before tax}}{\text{Turnover}}$
Quick Ratio	Financial structure	Percent	A measure of how well can the start-up meet its short-term financial liabilities	$\mathrm{QR} = \frac{\text{Current assets - Stocks}}{\text{Current liabilities}}$
Leverage Ratio	Financial structure	Percent	An indicator of how much capital comes in the form of debt compared to equity	$\mathrm{LR} = \frac{\text{Total debt}}{\text{Total equity}}$
Return on equity	Profitability	Percent	An indicator of how profitable a company is relative to its total equity	$\mathrm{ROE} = \frac{\text{Profit/Loss before tax}}{\text{Total equity}}$
Pre-tax return on invested capital	Profitability	Percent	An indicator of the start-up efficiency at allocating its capital to profitable investments	$\mathrm{ROIC} = \frac{\text{Profit/Loss before tax}}{\text{Current assets} + \text{Fixed Assets - Current liabilities - Cash}}$
Cashflow to capital expenditure ratio	Profitability/Financial structure	Percent	A measure of the start-up capability to acquire long-term assets using free cash flow	$\text{CF-CAPEX} = \frac{\text{P/L after tax} + \text{extr./other profit} + \text{depreciation}}{\text{Fixed assets}}$
Weighted average cost of capital	Profitability/cost of capital	Percent	An estimation of the start-up cost of capital, where each category of capital is proportionately weighted; based on Lünnemann and Mathä (2001)	$\mathrm{WACC} = \left(\frac{\text{Tot. debt}}{\text{Tot. assets}} * \frac{\text{Fin. expenses}}{\text{Tot. debt}}\right) + \left(\frac{\text{Tot. equity}}{\text{Tot. assets}} * \left(\text{LT rate}^{\dagger} + 5\%\right)\right)$

† Long-term debt rate. To approximate this value, this work uses historical data on 10-year government bond interest rates from ECB and OECD.

(continued)

Financial Indicator	Indicator dimension	Unit	Description	Formula
Cash to current assets ratio	Financial structure	Percent	An indicator of start-up liquidity	$CCA = \frac{\text{Cash}}{\text{Current assets}}$
Solvency ratio	Financial structure	Percent	A measure of how the start-up can meet its debt and other obligations	$SR = \frac{\text{Total equity}}{\text{Total assets}}$

Source: Orbis, author

D Factors affecting descriptive statistics of firm growth

This section explores two potential factors affecting the descriptive statistics of section 4.1 and 4.2, namely *selection bias* and *outliers*. Moreover, it portrays the effects of certain growth determinants outlined by Coad (2007): geographic, sectoral, macroeconomic and age-related characteristics of start-ups. On selection bias, Table D1 lists sample sizes for the analysed variables and horizons.

Table D1: *Usable* sample and non-missing rate by post-investment year

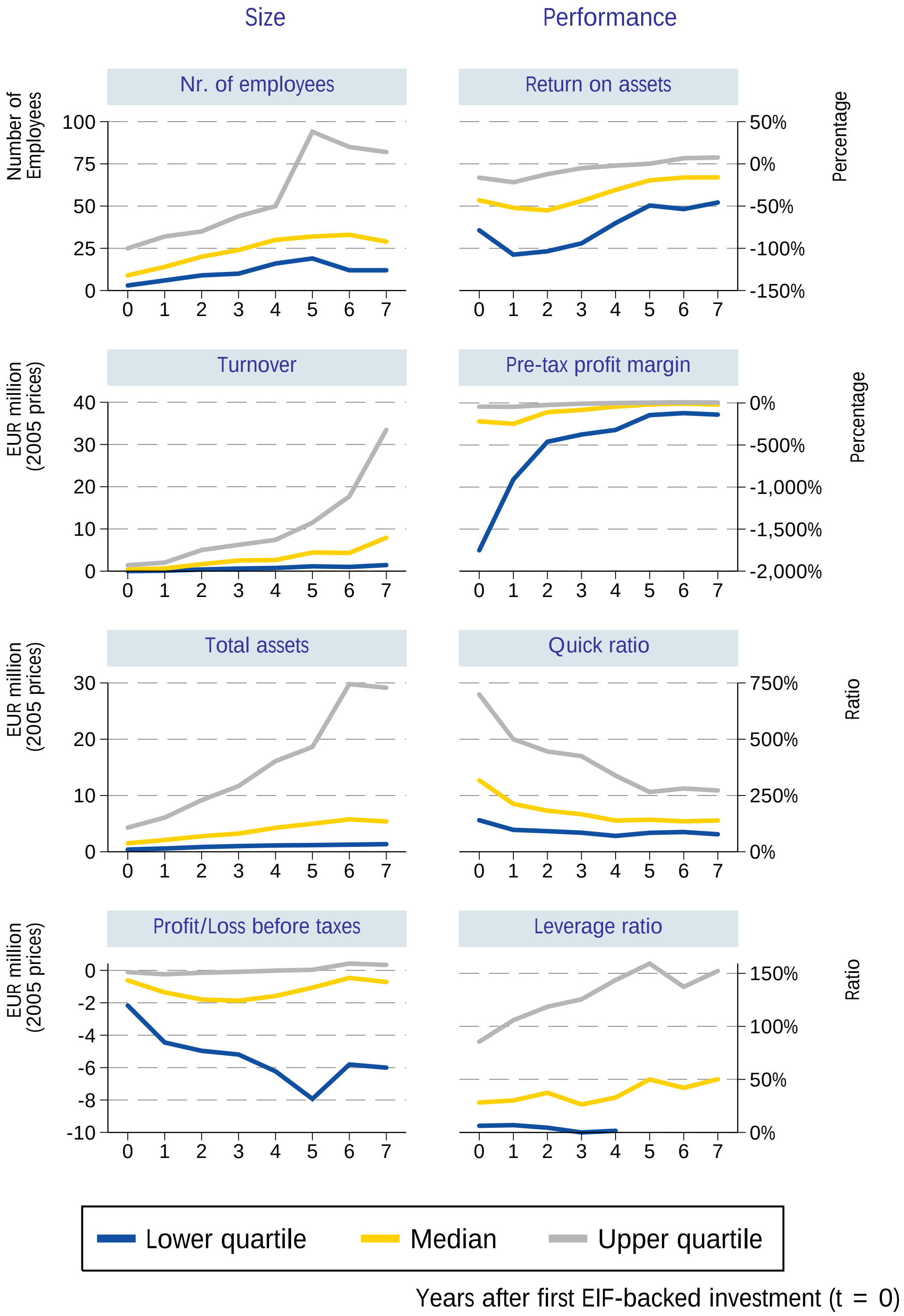
	Years after first EIF-backed investment (t = 0)							
Indicator	0	1	2	3	4	5	6	7
Economic size:								
Nr. of employees	938 (79.5%)	938 (82.6%)	938 (72.8%)	903 (70.6%)	809 (73.1%)	711 (80.8%)	656 (91.9%)	603 (95.5%)
Turnover	949 (87.4%)	949 (89.1%)	949 (78.2%)	932 (75.2%)	841 (82.0%)	756 (85.3%)	702 (89.7%)	655 (91.9%)
Total assets	1,375 (93.9%)	1,375 (92.5%)	1,375 (79.2%)	1,344 (74.5%)	1,227 (75.5%)	1,111 (76.7%)	1,029 (80.6%)	954 (81.2%)
Profit/Loss before taxes	962 (90.5%)	962 (90.2%)	962 (79.0%)	948 (76.1%)	863 (79.3%)	786 (82.6%)	740 (85.9%)	700 (86.2%)
Profitability and fin. structure:								
Return on assets	960 (90.3%)	960 (90.1%)	960 (78.7%)	946 (76.2%)	861 (79.0%)	784 (82.1%)	738 (85.7%)	698 (85.6%)
Pre-tax profit margin	835 (86.3%)	835 (87.6%)	835 (78.3%)	823 (76.3%)	749 (82.2%)	682 (85.9%)	641 (88.4%)	603 (88.7%)
Quick ratio	1,313 (84.6%)	1,313 (91.5%)	1,313 (77.6%)	1,205 (76.1%)	1,102 (75.7%)	1,012 (75.5%)	947 (77.4%)	883 (76.7%)
Leverage ratio	1,346 (94.5%)	1,346 (92.5%)	1,346 (78.9%)	1,316 (73.8%)	1,200 (75.1%)	1,089 (75.6%)	1,009 (79.2%)	936 (79.5%)
Additional indicators:								
Return on equity	776 (95.4%)	776 (86.9%)	776 (72.4%)	768 (67.4%)	700 (71.7%)	637 (73.9%)	602 (74.0%)	566 (76.1%)
Cashflow to capital expenditure ratio	753 (75.9%)	753 (83.6%)	753 (74.9%)	741 (74.3%)	682 (80.0%)	627 (86.2%)	589 (92.5%)	558 (92.8%)
Cash to current assets ratio	1,298 (91.3%)	1,298 (90.8%)	1,298 (78.2%)	1,273 (72.4%)	1,165 (73.3%)	1,058 (72.8%)	981 (76.9%)	906 (77.5%)
Solvency ratio	1,277 (95.9%)	1,277 (91.5%)	1,277 (75.7%)	1,254 (69.8%)	1,153 (68.4%)	1,046 (69.3%)	969 (72.5%)	896 (73.1%)
Weighted average cost of capital	717 (86.4%)	717 (87.8%)	717 (79.3%)	708 (76.4%)	646 (79.5%)	599 (81.3%)	568 (83.9%)	533 (85.5%)
Pre-tax return on invested capital	802 (87.0%)	802 (88.0%)	802 (78.9%)	792 (75.0%)	723 (78.9%)	663 (83.2%)	625 (86.2%)	589 (83.8%)
Total portfolio size	2,951	2,951	2,951	2,582	2,359	2,171	2,046	1,928

Note: each cell reports the number of *usable* companies for the given variable and the specific post-investment period. For a definition of *usable* company refer to Appendix B. Percentages in brackets represent the proportion of *usable* companies with actual non-missing data in the reference year.

Table D1 documents how, despite the significant effect of missing data, sample sizes tend to be large enough for the elaboration of descriptive statistics. Low rates of missing data support the view that residual sample attrition, however distributed, can only have a limited impact on the estimates.

Concerning the impact of outliers, Figure D1 portrays growth trends for the three main distribution quartiles. In addition to the median, these indicate the development of financial indicators at the lower and upper quartile (25th and 75th percentile respectively). Figure D1 corroborates the notion that patterns of start-up growth prove to be extremely diverse, with the consequence that even median values may fail to truly represent the "typical" start-up evolution path. An additional source of

Figure D1: Quartile growth trends of size and profit indicators



Note: the figure above portrays the heterogeneity of growth trends following an EIF-backed VC investment. Left-side charts show size trends, while right-side charts show performance and profitability trends. The *x-axis* counts the periods (in years) following the VC investments, where period 0 is the investment year. Statistics are computed using *response propensity weights* (Little, 1986). Methodological details are discussed in Appendix B.

heterogeneity are outliers. Outliers are a distinguishing feature of start-up growth, as evidenced in numerous studies (*e.g.* Calvino *et al*. 2015; Garnsey and Heffernan 2005).

The remainder of this appendix uses descriptive trends to investigate some possible determinants of start-up growth trends. To assess whether the factors mentioned above are linked with different median evolution trends, the analysis breaks down the portfolio of EIF-backed VC start-ups across these relevant dimensions. Statistics are thus computed using the approach described in section 3, albeit with a reduced set of auxiliary variables given the breakdown-induced smaller samples.[30]

This appendix collects the 85 different growth trends yielded by such analysis. The various patterns of start-up growth are organised as follows: each row portrays the development of a specific variable (employees, turnover, total assets, ROA, leverage ratio), while each column defines the subset of EIF-backed startups on which statistics were computed.

Figures D2 to D4 allow to glance through the different components that shape the aggregate trends discussed in section 4. First, differentiating across the three major investment periods — 1996 to 2001 (*pre*-dotcom), 2002 to 2007 (*post*-dotcom and *pre*-crisis) and 2008 to 2014 (*post*-crisis) — no particular insight seems to emerge. One exception is perhaps the significantly smaller size of post-crisis cohorts, compared to 2002-2007 investments, that can be observed in the first year in terms of assets and employment. However, these differences are quickly levelled out in the post-investment period or otherwise lose their significance.

Second, trends in employee size evidence the significantly steeper growth of start-ups based in UK and Ireland. However, the growth of the typical firm there seems to halt around the third year, while in other regions median employee growth is more stable.[31] Second, total assets at investment date are typically higher for France and Benelux than in other regions. These start-ups also typically show a higher level of indebtedness than their foreign peers (a feature shared also by Scandinavian companies). Slight differences emerge in terms of turnover growth and profitability trends, but overall regional breakdown is not particularly effective at disentangling the mechanism of start-up growth, at least not void of additional heterogeneity and uncertainty.

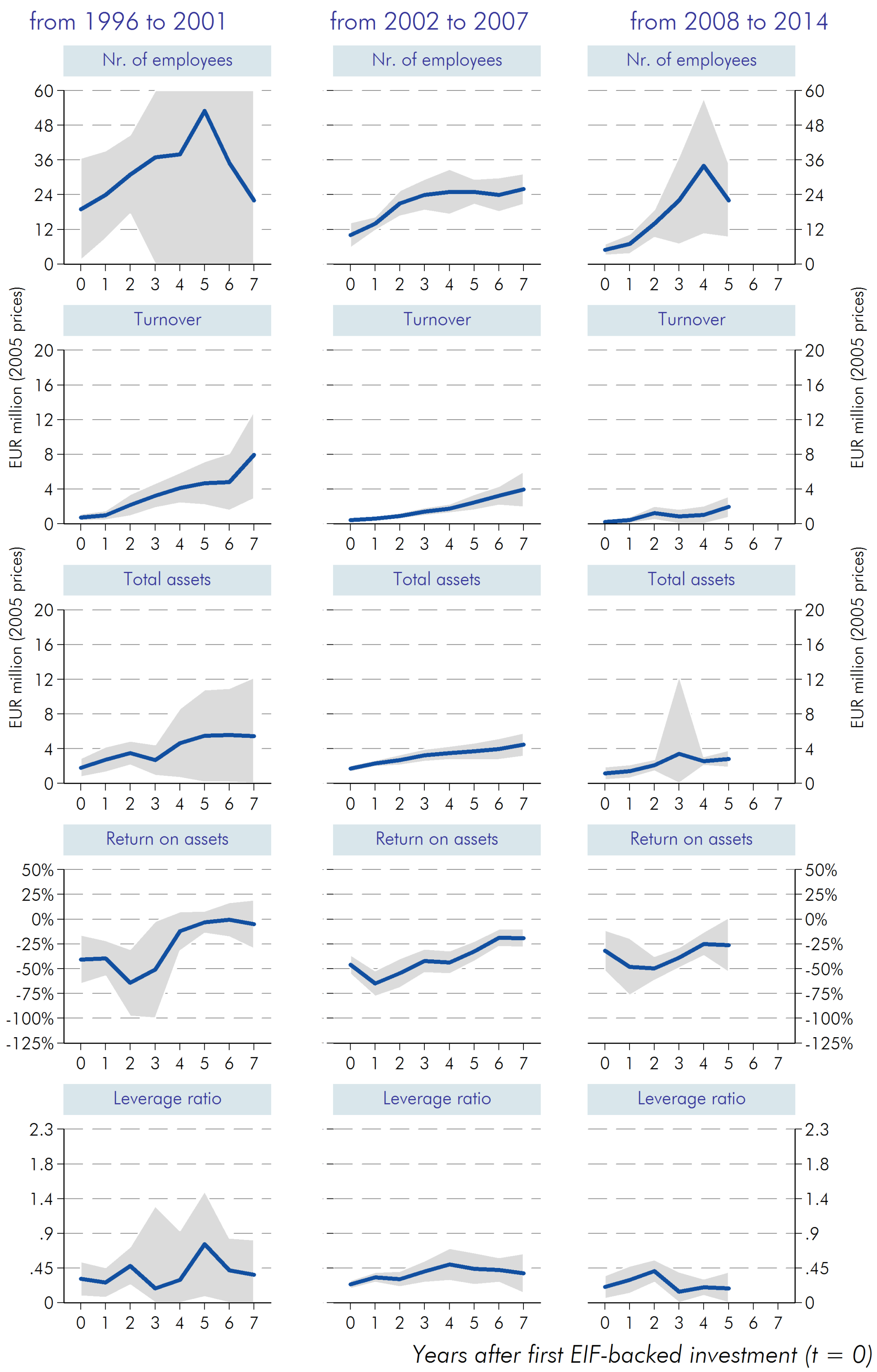
Lastly, sectoral disaggregation introduces a finer level of detail into the identification of different growth patterns. While the median ICT start-up does not have a significantly smaller employment footprint than its life science counterpart, growth rates for the former are typically higher. Faster growth slopes are observable also in terms of turnover and profitability, while asset-wise life science typically start bigger (but fail to keep this primacy thereafter).

---

[30] Additional details about the re-weigthing process included in Appendix B. Since the multitude of regression models employed to estimate statistical weights — one different model per subset and per indicator — is too sizeable to allow full reporting, some indications of the employed variables are provided instead.

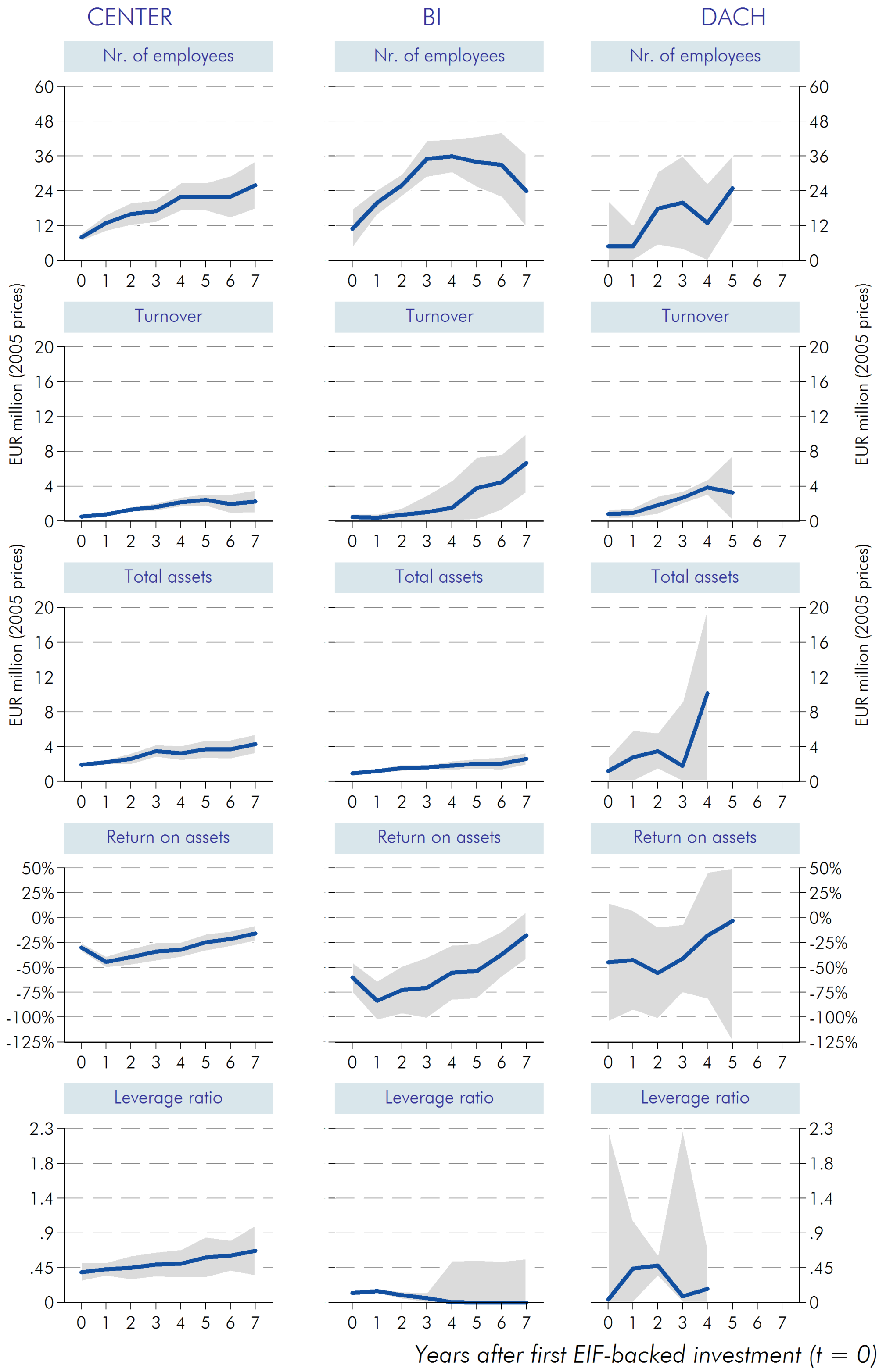
[31] There is not enough evidence to assert whether British- and Irish-based start-ups face a declining employee median size in the 5th post-investment period, as survivorship bias is particularly strong in this region.

Figure D2: Evolution of EIF-backed start-ups by investment period



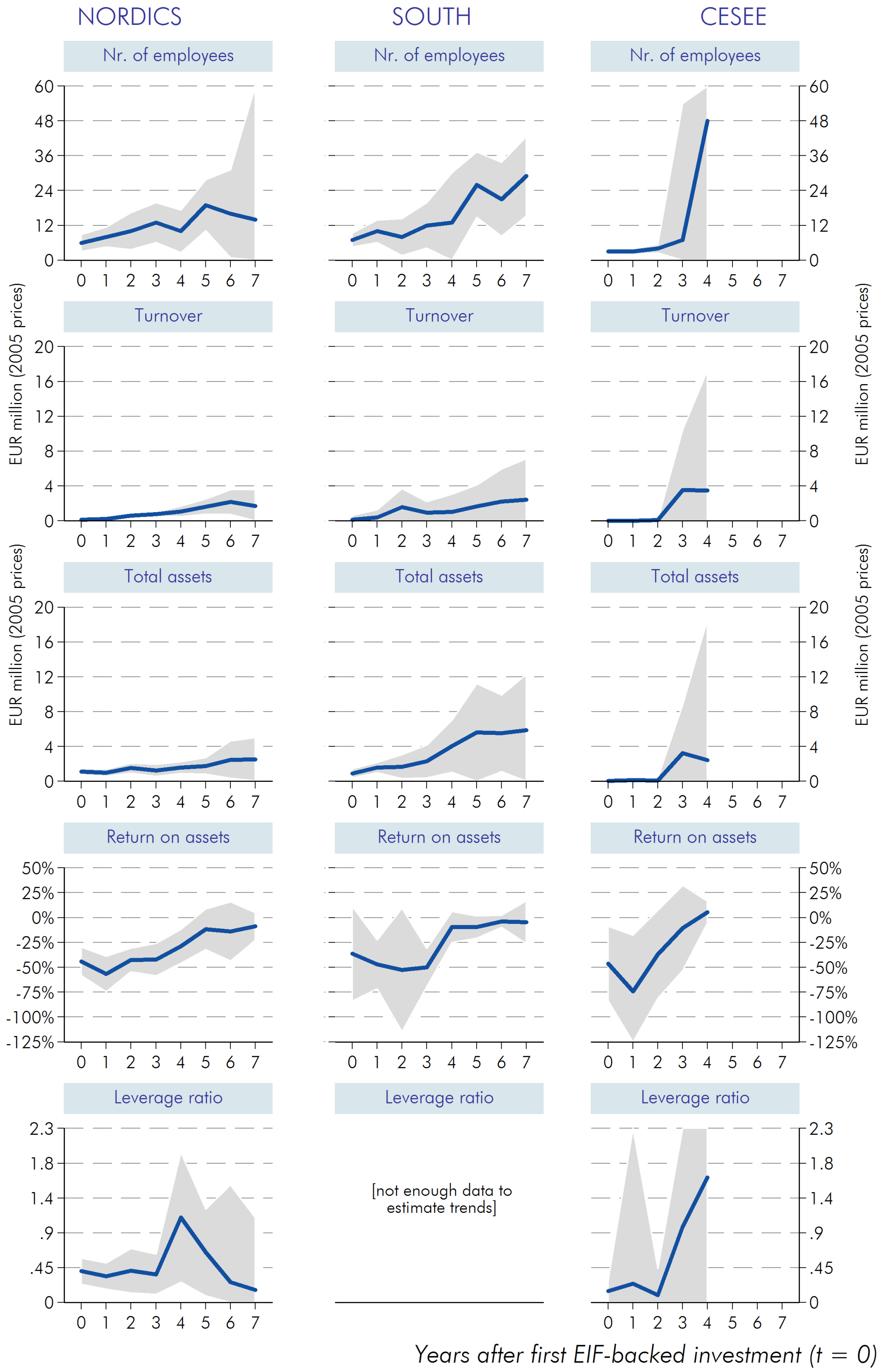
Note: the figure above portrays the development of financial indicator following an EIF-backed VC investment, across a wide range of subgroups. All charts show median values, complemented with 95% confidence intervals. The *x-axis* counts the periods (in years) following the VC investments, where period 0 is the investment year. Statistics are computed using *response propensity weights* (Little, 1986). Methodological details are discussed in Appendix B.

Figure D3: Evolution of EIF-backed start-ups by geographic area



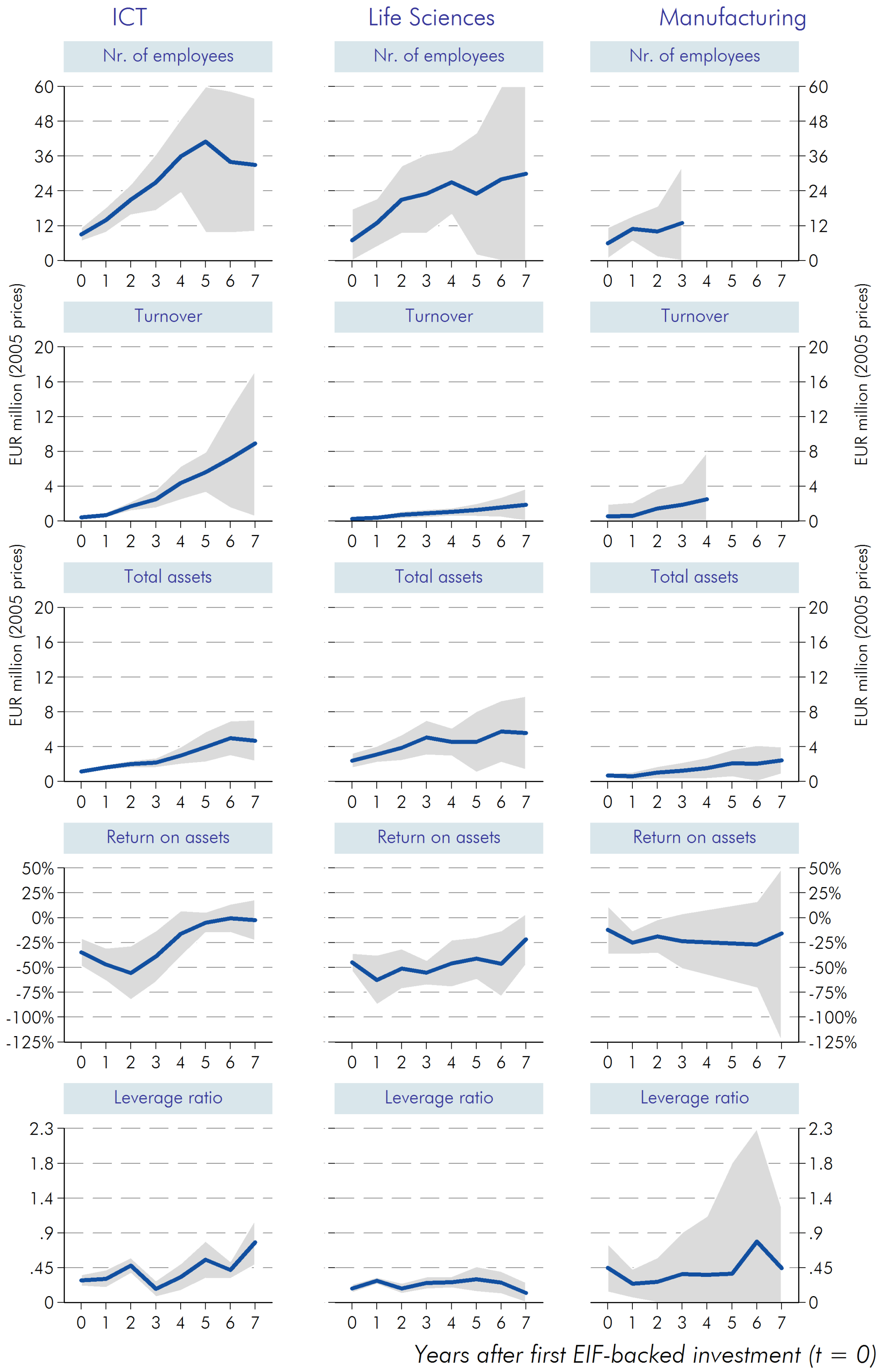
Note: the figure above portrays the development of financial indicator following an EIF-backed VC investment, across a wide range of subgroups. All charts show median values, complemented with 95% confidence intervals. The *x-axis* counts the periods (in years) following the VC investments, where period 0 is the investment year. Statistics are computed using *response propensity weights* (Little, 1986). Methodological details are discussed in Appendix B.

(Figure D3 continued)



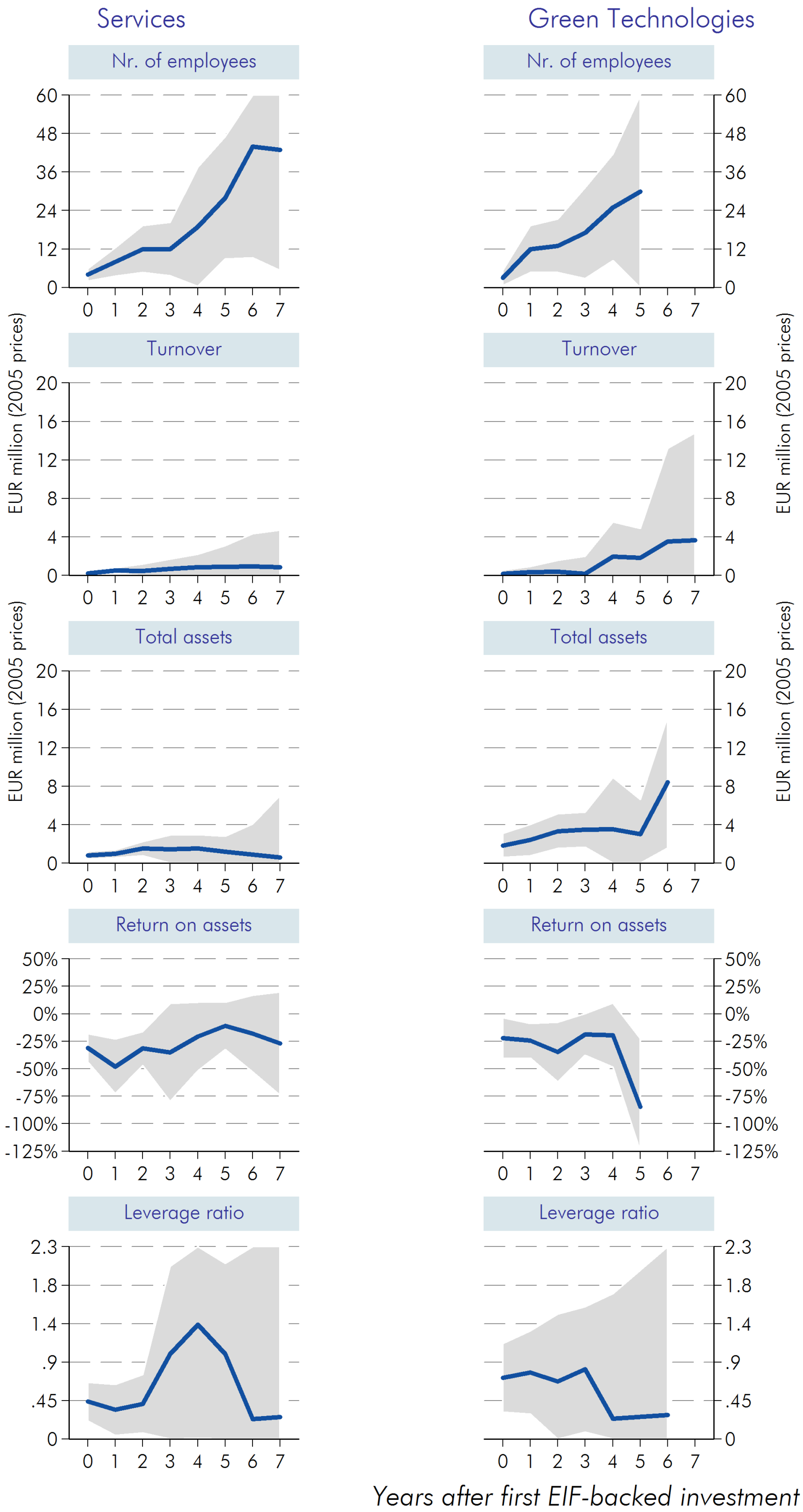
Note: the figure above portrays the development of financial indicator following an EIF-backed VC investment, across a wide range of subgroups. All charts show median values, complemented with 95% confidence intervals. The *x-axis* counts the periods (in years) following the VC investments, where period 0 is the investment year. Statistics are computed using *response propensity weights* (Little, 1986). Methodological details are discussed in Appendix B.

Figure D4: Evolution of EIF-backed start-ups by macro-industry



Note: the figure above portrays the development of financial indicator following an EIF-backed VC investment, across a wide range of subgroups. All charts show median values, complemented with 95% confidence intervals. The *x-axis* counts the periods (in years) following the VC investments, where period 0 is the investment year. Statistics are computed using *response propensity weights* (Little, 1986). Methodological details are discussed in Appendix B.

(Figure D4 continued)



Note: the figure above portrays the development of financial indicator following an EIF-backed VC investment, across a wide range of subgroups. All charts show median values, complemented with 95% confidence intervals. The *x-axis* counts the periods (in years) following the VC investments, where period 0 is the investment year. Statistics are computed using *response propensity weights* (Little, 1986). Methodological details are discussed in Appendix B.

E Additional descriptive statistics

The set of additional metrics begins with return-on-equity (ROE) and pre-tax return on invested capital (pre-tax ROIC). Similarly to ROA, these ratios assess the start-up's ability to efficiently allocate capital to generate profits: where ROE may be considered bearing the perspective of the investor, ROIC instead portrays the views of the entrepreneur. A further indicator linked with the assessment of firm profitability is the weighted average cost of capital (WACC). Unlike other measures, WACC cannot be directly derived from start-up financials, primarily for lack of data on the cost of equity: it is thus estimated following the approach of Lünnemann and Mathä (2001). The weighted average cost of capital can be compared against ROIC to assess whether the company's use of capital is more profitable than its cost. However, for the reasons expressed above, such comparison may not be particular insightful as it concerns the first years, typically unprofitable, of start-up activity. Nevertheless, the estimated WACC is a valuable tool to assess whether start-up development is characterised by higher costs at entry (*i.e.* investment year) which further decrease upon observed company growth.

The remaining three indicators — cashflow to capital expenditure ratio, cash to current assets ratio and solvency ratio — contribute with different angles on the financial structure of start-ups. Cashflow over capital expenditure indicates the extent to which a company is able to acquire long-term assets with its excess cash. Complementing the quick ratio, cash over current assets signals how liquid are start-up assets. Finally, the solvency ratio (defined as equity over total assets) expresses the start-up's ability to meet its debts and other obligations. Evolution trends of these additional indicators are depicted in Figure E1.

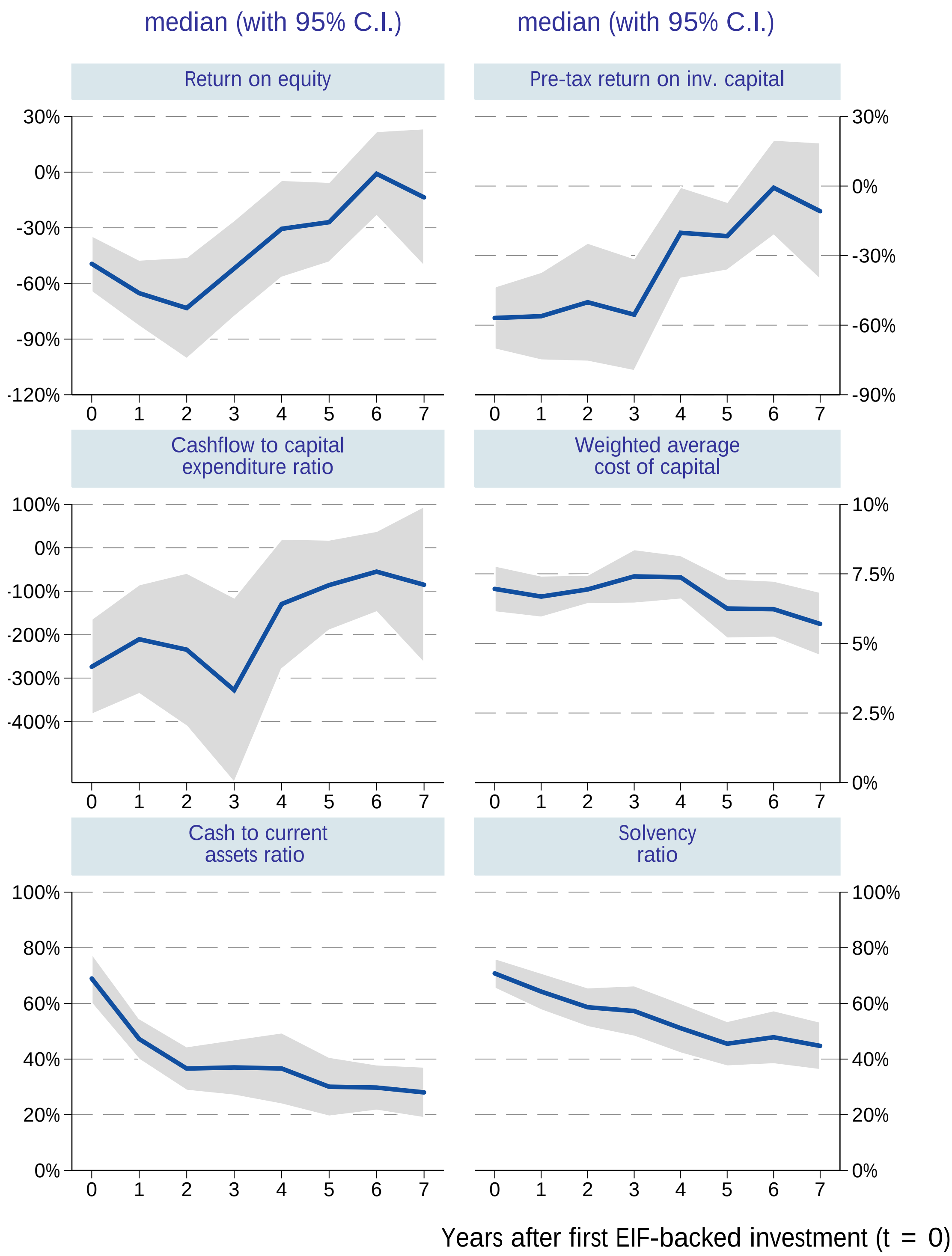
Drawing on the findings of section 4.1 and 4.2, the additional indicators described above provide a further confirmation in support of the arguments explored thus far. The two profitability ratios (ROE and ROIC) follow a trend that appears quite similar to the ROA of start-ups: all median post-investment profitability trends face a clear inversion from the second year onward, with ROIC facing a further one-year lag.[32] The analysis of the WACC shows that the median EIF-backed VC start-up faces a cost of capital of around 7.5%. This persists up until the fifth post-investment year, then it quickly converges to 5%.[33] As per the remaining indicators, cashflow over capital expenditure shows that most start-ups would be unable to sustain long-term investments with the liquid assets generated by their day-to-day activity.[34] Cash over current assets trends confirm the view that start-ups are predominantly endowed with liquid assets, which they consume at a relatively fast pace. Lastly, solvency ratio offers a similar evolution to the leverage ratio, albeit from the opposite perspective. The findings are consistent with Robb and Robinson (2014), who find at investment date that "about 40% of [start-up] total capital is financed through outside debt" (*ibid*, p. 167).

[32] Time-wise, this phenomenon occurs when the median start-up age is 3–4 years.

[33] Note: the comparison of these measures against other publicly available sources would be flawed: as estimated values are strongly influenced by the assumptions set forth in Lünnemann and Mathä (2001), the only practical evaluation approach would be to estimate the same figure for a comparison group of companies. However, such aspect is beyond the scope of this work.

[34] However, it could also be that VC investments cause start-ups to adopt a less conservative behaviour towards their cashflow use. Additional research is needed to shed further light on such finding.

Figure E1: Average and median growth trends of additional indicators



Note: the figure above portrays additional financial ratios following an EIF-backed VC investment. All reported values show median trends. The *x-axis* counts the periods (in years) following the VC investments, where period 0 is the investment year. Statistics are computed using *response propensity weights* (Little, 1986). Methodological details are discussed in Appendix B.

F Cluster analysis methods

This appendix compares the model-based cluster analysis approach with two alternative clustering techniques. In doing so, it provides a justification of the final number of clusters selected for the analysis, and discusses alternative clustering solutions. The three clustering methods are compared by means of a confusion matrix.

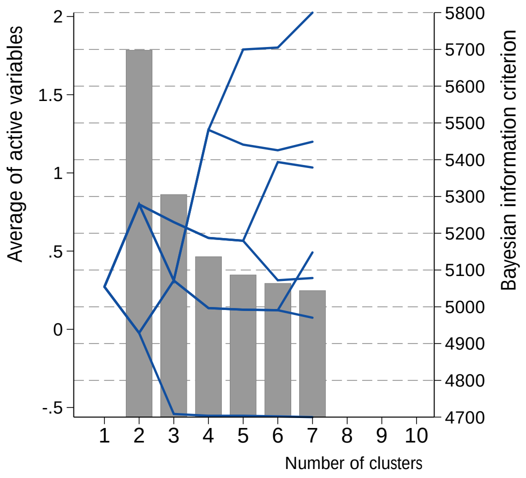
As described in section 5, the five clustering dimensions — the *active* variables — are pre-processed via the *neglog* transformation (Whittaker *et al.*, 2005) and further standardisation. The analysis excludes start-ups with incomplete data, addressing the shortcomings of complete-case analysis with the *response propensity stratification* approach discussed in Appendix B. Severe data loss is still observed for some clusters of data (typically pertaining to older investment vintages), particularly for baseline year data. For such reason, about half of the baseline data from BvD Orbis is integrated with information from fund reports, when available, and about 10% of baseline data is imputed following Welch *et al*. (2014). It is important to remark that integration and imputation of data only affects the indicators of total assets, nr. of employees, and only very marginally turnover. The approximation error for the former two variables is considered low, as these tend to be more stable in a neighbourhood of the baseline period.

The performance of the *latent class* cluster analysis is compared against hierarchical clustering, performed via Ward's method, and *k-means* clustering. The details of these alternative approaches will not be reviewed here: the reader is referred to Everitt (2011) for a comprehensive exposition. The comparison is carried by means of *clustergrams* (Schonlau, 2002). A clustergram is a visual representation of the clustering process. Values on the x-axis represent the number of groups resulting from each iteration of the clustering process, while figures on the y-axis portray a statistic (in this case, the average) of the clustering variables. While the statistic bears no practical interpretation, it gives an indication on the positioning of each cluster, hence the "distance" among clusters. Moreover, clusters are "connected" across further iterations, helping to track the creation of additional clusters of data. Results are presented in Figure F2. Choosing an optimal number of clusters typically entails the application of a "stopping rule" for the selected clustering method. In the case of Ward's and k-means technique, the Calinski-Harabasz pseudo-F is used, portrayed as grey bars. This statistic reflects the "outer balance" of clusters, *i.e.* how different they are among each other, and therefore a higher value indicates better performance. On the contrary, the model-based approach (Figure F2a) is evaluated on the basis of the Bayesian Information Criterion (BIC), hence smaller values are better.

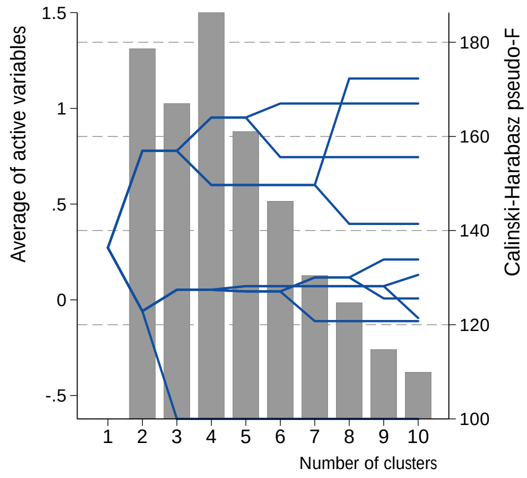
The prediction of the latent class cluster model does not consist of a unique cluster for each vector of active variables. This is a key difference with both Ward's and *k-means* method. Instead, it fits the data assuming a mixture model with n subgroups, where n is chosen a priori. Once a (locally) optimising set of parameters for the distribution is found, it is possible to use such parameters to estimate the vector posterior probability of affiliation to a given latent class. In other words, each combination of active variables is provided with n posterior probabilities, one for each selected class. Growth profiles are assumed mutually exclusive. For this reason, this exercise imposes that the cumulative sum of all n probabilities amounts to 1. Exceeding cumulative probabilities are rescaled accordingly. For each combination of active variables, cluster affiliation is selected according to the highest posterior probability observed among the n clusters.

The first aspect to clarify with regards to Figure F2 is that the model-based approach is unable to find solutions with more than 7 classes.[35] Moreover, while Ward's method and the k-means approach point out almost unequivocally to the 4 clusters solution, the model-based approach reveals that further clusters are in fact able to improve the likelihood of the model. However, there are two practical reasons for the final choice of 4 clusters. First, the improvement in terms of BIC signficantly decreases after the 4th cluster. Second, the additional growth profiles generated by using 5 and 6 clusters correspond to micro-clusters within the out-performing category (see Figure F2a), comprising a total of 5.5% portfolio companies.

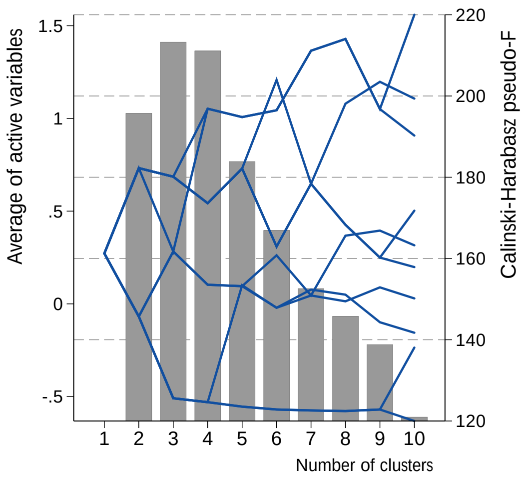
Figure F2: Clustergrams of different clustering methods



(a) Latent class approach



(b) Ward's method



(c) *K-means* method

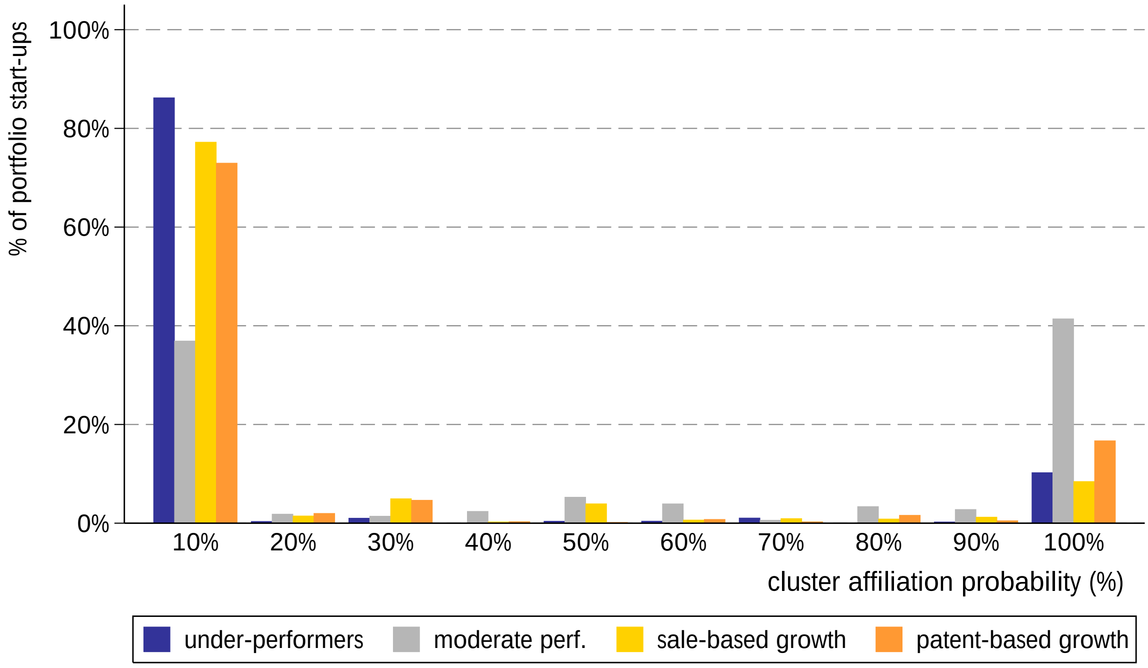
Only in the 7th cluster solution it is possible to "extract" a different subset from the moderately growing profile. The additional profile totals 12% of companies and contains a subset of moderately growers characterised by similar growth values in total assets, turnover and patenting growth, but a more

35 More precisely, a reasonably fast solution: the optimisation algorithm typically reaches a flat area of the log-likelihood function and is thus unable to converge. Several different sets of starting points have been tested with little to no avail.

pronounced growth in terms of employees, as well as a positive valuation CAGR. All in all, the additional solutions do not seem to introduce significant information, while at the same time they certainly render the exposition of results more cumbersome.

Figure F3 depicts the distribution of the estimated cluster posterior probabilities. As expected, probabilities tend to be significantly polarised: this signals the success of the latent class model to identify well-distanced classes. The least polarised groups are moderate performers and sale-based growers.

Figure F3: Decile distribution of cluster posterior probability



Note: deciles computed using weights. N = 1,846; n = 440.

Table F1 is a confusion matrix comparing the three clustering methods. Clustering profiles typically coincide among the three classification strategies. However, the model-based approach appears more conservative against out-performers, particularly for sale-based growers.

Table F1: Confusion matrix of clustering methods

	Under-performers		Moderate growers		Sales-based growers		Patent-based growers	
Latent class profile	Ward's	*k-means*	Ward's	*k-means*	Ward's	*k-means*	Ward's	*k-means*
under-performers	100%	91.81%	6.78%	0%	0%	0%	0%	0%
moderate performers	0%	8.18%	91.25%	100%	43.93%	48.26%	23.9%	9.6%
sale-based growth	0%	0%	0%	0%	54.14%	51.73%	1.62%	0%
patent-based growth	0%	0%	1.96%	0%	1.92%	0%	74.47%	90.39%

Notes: Statistics computed using weights. N = 440.

In conclusion, the latent class model is observed to perform well with a 4-class solution. The results are most similar to the solutions provided by alternative approaches, *i.e.* Ward's method and k-means, although the latent class model is more conservative about out-performing profiles. The latent class model is thus selected as the benchmark model in this paper.

G Cluster comparison: determinants of growth profiles

Variable	Categories	Total sample	Usable sample†	N	Underperformers %	Underperformers Z	Moderate growth %	Moderate growth Z	Sales-based growth %	Sales-based growth Z	Patent-based growth %	Patent-based growth Z
Cluster distribution					12.40%		55.46%		12.14%		20.00%	
Startup's macro-region location	CENTER	27.36%	29.35%		29.00%	-.12	30.80%	1.5	15.89%	-4.63***	33.71%	2.02**
	BI	28.55%	29.04%		33.72%	1.63	24.55%	-4.65***	39.14%	3.48***	32.47%	1.59
	DACH	21.89%	20.69%		14.13%	-2.56**	20.43%	-.3	27.22%	2.52**	21.52%	.43
	NORDICS	11.16%	10.77%		11.88%	.56	12.09%	2**	4.67%	-3.08***	10.13%	-.43
	SOUTH	9.86%	9.12%		11.28%	1.19	10.84%	2.8***	10.54%	.77	2.17%	-5.09***
	CESEE	1.19%	1.02%	1846	0.00%		1.29%	1.23	2.55%	2.37**	0.00%	
Startup's macro-industry	ICT	61.32%	62.93%		71.11%	2.68***	70.12%	6.99***	58.73%	-1.36	40.49%	-9.79***
	Life Sciences	25.68%	25.42%		14.99%	-3.8***	16.45%	-9.68***	35.91%	3.77***	50.41%	12.09***
	Services	7.31%	5.77%		7.31%	1.05	7.84%	4.19***	2.54%	-2.16**	1.01%	-4.3***
	Manufacturing	4.23%	4.15%		2.49%	-1.32	4.64%	1.16	2.82%	-1.04	4.61%	.49
	Green Technologies	1.46%	1.73%	1846	4.12%	2.9***	0.94%	-2.82***	0.00%		3.47%	2.81***
Startup's vintage period	1996-2001	46.75%	44.16%		46.82%	.85	43.96%	-.19	42.90%	-.4	43.83%	-.14
	2002-2007	37.92%	41.49%		34.25%	-2.33**	42.15%	.63	38.78%	-.86	45.79%	1.84*
	2008-2014	15.33%	14.35%	1846	18.93%	2.07**	13.89%	-.61	18.31%	1.77*	10.38%	-2.39**
Age at (first) inv.	0-2	63.22%	62.48%		73.62%	3.65***	53.29%	-8.91***	66.11%	1.17	78.83%	7.12***
	2-5	28.22%	28.45%		21.81%	-2.33**	34.80%	6.61***	25.61%	-.99	16.69%	-5.49***
	5-10	8.56%	9.07%	1846	4.58%	-2.48**	11.90%	4.63***	8.28%	-.43	4.48%	-3.37***
Startup's exit route	Write-Off	29.31%	30.74%		75.35%	15.32***	25.85%	-4.98***	17.91%	-4.35***	24.42%	-2.89***
	Trade sale (MOC >= 1)	14.08%	14.74%		4.20%	-4.71***	16.41%	2.21**	21.43%	2.95***	12.59%	-1.28
	IPO	6.61%	7.29%		0.00%		5.95%	-2.42**	12.78%	3.3***	12.19%	3.97***
	"Trade sale" "(MOC < 1)"	27.95%	27.14%	1439	15.97%	-3.98***	33.06%	6.25***	24.58%	-.9	19.19%	-3.76***

* $p<0.10$, ** $p<0.05$, *** $p<0.01$

Notes: the table shows the distributional differences across growth clusters. Row 1 contains the proportion of each growth cluster in the overall portfolio. Column 1 indicates the reference passive variable, and column 2 its classes of distribution. Column 3 shows the total portfolio distribution, while Column 4 shows the sample distribution (N = 440) computed with *response propensity weights* (Little, 1986). Column 6 - 13 are to be read in pairs: each pair of columns portrays the variable distribution in the reference cluster (%) and the Z statistic resulting from a two-proportions test. Each test is carried by splitting the sample in two subgroups: a first composed by firms of the reference cluster, and a second with all other clusters.

H Out-of sample estimation of cluster affiliation propensity

This section discusses the model specification and estimation strategy of cluster probabilities. This exercise can be considered a form of imputation of missing values, albeit less formalised. This approach is preferred despite the possibility to use the *response propensity weights* discussed in Appendix B, which are based on more solid theoretical grounds. There is one major reason for this choice: the significant missing rate (about 76%) makes it impossible to generate a representative geographical distribution of the EIF VC portfolio from sample data.

A second-best approach is thus to impute for each portfolio company cluster affiliation propensities, based on a set of explanatory variables that are observable for the entire portfolio. The principle is similar to the method described in Appendix B. Conditional on $\mathbf{a}$ (a set of start-up, investor and investment attributes), cluster affiliation probability $\mathrm{p_c}$ is assumed independent of the response mechanism r. It thus possible to regress $\mathrm{p_c}$ on $\mathbf{a}$ and obtain $\hat{\mathrm{p}}_\mathrm{c}$, the predicted probability of cluster affiliation. This approach allows to take advantage of the posterior probabilities estimated by the latent class model, and can be consider more fine-grained approach than *e.g.* estimating propensities via a probit or logit model on cluster indicators.

Given the bounded nature of $\mathrm{p_c} \in [0, 1]$, an appropriate method to estimate $\mathrm{p_c}$ is the generalised linear model (GLM) with binomial family and probit link. This approach is practically equivalent to the *fractional logit* model of Papke and Wooldridge (1996), and it overcomes the drawbacks of conventional least squares estimation when the dependent variable is bounded between 0 and 1. However, as described in Appendix B observable data suffers from sample selection bias, due to the two stages of response labelled *matched* and *usable*. To address this issue, the two-step Heckman estimator is used to account for selection on unobservables. Following Heckman (1979), in the first step the selection equation is estimated by regressing the *matched* indicator on a set of variables $\mathbf{b}$. The inverse Mill's ratio is thus calculated on the basis of $\hat{\mathrm{m}} = \mathrm{pr}\left(\mathrm{m} = 1|\mathbf{b}\right)$, the predicted probability of m. The inverse Mill's ratio is thus used in the outcome regression, estimated via GLM. Dubin and Rivers (1989) discuss how Heckman's correction produces biased results for nonlinear second step regressions. The authors discuss a correct implementation of the Heckman two-step process for the case of binary choice models, but no implementation could be retrieved for the case of generalised linear models. Moreover, by implementing a second stage OLS regression, results are not observed to change substantially.[36] For this reason, predicted cluster affiliation probabilities $\mathrm{p_c}$ are estimated via a two-step Heckman selection model with a second stage GLM estimation.

Results are portrayed in Table H2. Column (1) portrays the results of the first stage probit regression, where the dependent variable is equal to 1 for matched companies and 0 otherwise. Columns (2) to (5) portray the results of the second stage GLM regression, estimated using standard errors robust to heteroskedasticity. The *exclusion restriction* for the selection equation is accounted for with the use of additional variables. Consistent with Appendix B, these pertain to the nature of investors (*e.g.* the fund sectoral focus) as well as details on the collection of accounts performed by Bureau Van Dijk (*e.g.* the type of national institution that collects financial records).

To derive the geographical distribution of growth patterns (Figure 8), the estimated start-up cluster probabilities $\hat{\mathrm{p}}_\mathrm{c}$ are aggregated at the city-level, taking the average value as the representative

[36] Let alone the fact that predicted probabilities will not be bounded to the $[0, 1]$ interval anymore.

statistic.

Table H2: Predictive model for cluster affiliation propensity.

Variable	Matched eq. (1)	under-perf. (2)	moderate (3)	sale-based (4)	patent-based (5)
	PROBIT	GLM	GLM	GLM	GLM
Investment features:					
Investment year	0.104***				
	(0.0189)				
Nr. of investment rounds		-0.257***	-0.061	0.109*	0.153**
		(0.0690)	(0.0426)	(0.0489)	(0.0523)
Nr. of investor funds	0.228	-0.438*	-0.193	0.283	-0.014
	(0.1372)	(0.2185)	(0.1303)	(0.1462)	(0.1448)
Nr. of quarters in the EIF portfolio	0.034***				
	(0.0044)				
Start-up has high innovation potential†	0.056				
	(0.1767)				
Investee features:					
Age at investment	0.404***				
	(0.0656)				
Age squared at investment	-0.040***				
	(0.0088)				
BvD collection:					
Company matched via BvD algorithm†		-0.026	-0.010	0.016	0.185
		(0.1941)	(0.1231)	(0.1452)	(0.1478)
Economic environment:					
Country PPI (log-form)		4.574**	-1.091	-0.538	-1.002
		(1.4624)	(1.0875)	(1.1856)	(1.2141)
Matching equation:					
Inverse Mill's ratio		1.531**	-1.929**	0.933	0.871
		(0.5124)	(0.6344)	(0.6496)	(0.5528)
Constant term	-210.000***	-0.399	0.538	-1.803***	-1.590***
	(38.0103)	(0.5894)	(0.3896)	(0.4135)	(0.4685)
Standardised Financial records × Dummy for small strata		Yes	Yes	Yes	Yes
Standardised Financial records × Start-up exit type		Yes	Yes	Yes	Yes
Standardised Financial records × Start-up investment period		Yes	Yes	Yes	Yes
Start-up investment period × Focus of VC fund		Yes	Yes	Yes	Yes
Focus of VC fund	Yes				
Institution storing financial records	Yes				
Macro-region of VC fund	Yes				
Start-up exit type	Yes				
Start-up investment period		Yes	Yes	Yes	Yes
Start-up macro-region		Yes	Yes	Yes	Yes
Start-up stage at investment		Yes	Yes	Yes	Yes
Pseudo R-squared	0.27	0.35	0.12	0.14	0.18
N	1846	440	440	440	440

* $p<0.05$, ** $p<0.01$, *** $p<0.001$; † dichotomic variable.

Using the software in Pisati (2007), the geographical distribution of cluster probabilities is further estimated via quartic kernel function and a fixed bandwidth or approx. 17,500 km^2, *i.e.* the area of a circle with diameter 150km. Specifically, the entire area of Europe[37] is split into 50,705 evenly distributed and non-overlapping hexagonal areas (called *unit areas*). For each unit area s, the quantity $\lambda_{s,c}$ is computed as the probability density function of cluster c in area s. The quantity $\lambda_{s,c}$ intuitively defines the expected number of start-ups with growth profile c in area s. The cluster affiliation probability in each small area $p_{c,s}$ is thus calculated as follows

$$p_{c,s} = \frac{\lambda_{s,c}}{\sum_{c=1}^{4} \lambda_{s,c}}$$

i.e. the expected proportion of start-ups with growth profile c in area s.

Figure 9 in section 6.4.1 portrays an index based on the affiliation probabilities for the two out-performing growth profiles. To simplify notation, consider $c = sale$ for sale-based growers, and $c = patent$ for patent-based growers. The index $\zeta_{sale,patent}$ is calculated as follows

$$\begin{aligned}\zeta_{s,(sale,patent)} &= \left(p_{sale,s} - p_{patent,s}\right) \times \left(\lambda_{sale,s} + \lambda_{patent,s}\right) \\ &= \left(\frac{\lambda_{sale,s} + \lambda_{patent,s}}{\sum_{c=1}^{4} \lambda_{s,c}}\right) \times \left(\lambda_{sale,s} - \lambda_{patent,s}\right)\end{aligned}$$

The index $\zeta_{s,(sale,patent)}$ weighs the surplus of sale- or patent-based growers by the relative proportion of out-performing profiles in area s. In this way, regions with a particular bias towards one of these two profiles but few out-performers overall will be considered more balanced, as the result may be due to a few occasional out-performers. Table H3 shows a number of descriptive statistics for the index mentioned above.

Table H3: Descriptive statistics for $\zeta_{sale,patent}$ and its components

variable	mean	median	min	max	p25	p75	N
$\lambda_{sale,s} - \lambda_{patent,s}$	-0.5953	0.0016	-60.6009	13.7302	-0.1878	0.1179	10,194
$\frac{\lambda_{sale,s} + \lambda_{patent,s}}{\sum_{c=1}^{4} \lambda_{s,c}}$	0.2703	0.2487	0.0466	0.9269	0.1673	0.3613	10,194
$\zeta_{sale,patent}$	-0.2115	0.0002	-17.0756	5.0260	-0.0544	0.0252	10,194

Notes: the descriptive statistics are computed on the (approximate) portion of European territory benefiting from EIF-backed investments in the 1996 – 2010 period, *i.e.* only if $\sum_{c=1}^{4} \lambda_{s,c} > 0$.

The distributions of $\zeta_{sale,patent}$ in Table H3 shows very long left and right tails, where the left tail is the longest. This reflects the distribution of the profile surplus $\lambda_{sale,s} - \lambda_{patent,s}$. Figure 9 gives a visual interpretation of the distribution at hand. The distribution classes are selected on the 10th, 25th, 50th, 75th and 90th percentile respectively.

[37] This includes all territories portrayed in Figure 8, hence not limited to EU28 countries.

About...

...the European Investment Fund

The European Investment Fund (EIF) is Europe's leading risk finance provider for small and medium-sized enterprises (SMEs) and midcaps, with a central mission to facilitate their access to finance. As part of the European Investment Bank (EIB) Group, EIF designs, promotes and implements equity and debt financial instruments which specifically target the needs of these market segments.

In this role, EIF fosters EU objectives in support of innovation, research and development, entrepreneurship, growth, and employment. EIF manages resources on behalf of the EIB, the European Commission, national and regional authorities and other third parties. EIF support to enterprises is provided through a wide range of selected financial intermediaries across Europe. Since its inception in 1994, EIF has supported over 1.8 million SMEs.

EIF is a public-private partnership whose tripartite shareholding structure includes the EIB, the European Union represented by the European Commission and various public and private financial institutions from European Union Member States and Turkey. For further information, please visit www.eif.org.

...EIF's Research & Market Analysis

Research & Market Analysis (RMA) supports EIF's strategic decision-making, product development and mandate management processes through applied research and market analyses. RMA works as internal advisor, participates in international fora and maintains liaison with many organisations and institutions.

...EIF Working Papers

The EIF Working Papers are designed to make available to a wider readership selected topics and studies in relation to EIF's business. The Working Papers are edited by EIF's Research & Market Analysis and are typically authored or co-authored by EIF staff, or written in cooperation with EIF. The Working Papers are usually available only in English and typically distributed in electronic form (pdf).

EIF Working Papers

2009/001	Microfinance in Europe – A market overview. November 2009.
2009/002	Financing Technology Transfer. December 2009.
2010/003	Private Equity Market in Europe – Rise of a new cycle or tail of the recession? February 2010.
2010/004	Private Equity and Venture Capital Indicators – A research of EU27 Private Equity and Venture Capital Markets. April 2010.
2010/005	Private Equity Market Outlook. May 2010.
2010/006	Drivers of Private Equity Investment activity. Are Buyout and Venture investors really so different? August 2010
2010/007	SME Loan Securitisation – an important tool to support European SME lending. October 2010.
2010/008	Impact of Legislation on Credit Risk – How different are the U.K. and Germany? November 2010.
2011/009	The performance and prospects of European Venture Capital. May 2011.
2011/010	European Small Business Finance Outlook. June 2011.
2011/011	Business Angels in Germany. EIF's initiative to support the non-institutional financing market. November 2011.
2011/012	European Small Business Finance Outlook 2/2011. December 2011.
2012/013	Progress for microfinance in Europe. January 2012.
2012/014	European Small Business Finance Outlook. May 2012.
2012/015	The importance of leasing for SME finance. August 2012.
2012/016	European Small Business Finance Outlook. December 2012.
2013/017	Forecasting distress in European SME portfolios. May 2013.
2013/018	European Small Business Finance Outlook. June 2013.

2013/019	SME loan securitisation 2.0 – Market assessment and policy options. October 2013.
2013/020	European Small Business Finance Outlook. December 2013.
2014/021	Financing the mobility of students in European higher education. January 2014.
2014/022	Guidelines for SME Access to Finance Market Assessments. April 2014.
2014/023	Pricing Default Risk: the Good, the Bad, and the Anomaly. June 2014.
2014/024	European Small Business Finance Outlook. June 2014.
2014/025	Institutional non-bank lending and the role of debt funds. October 2014.
2014/026	European Small Business Finance Outlook. December 2014.
2015/027	Bridging the university funding gap: determinants and consequences of university seed funds and proof-of-concept Programs in Europe. May 2015.
2015/028	European Small Business Finance Outlook. June 2015.
2015/029	The Economic Impact of EU Guarantees on Credit to SMEs - Evidence from CESEE Countries. July 2015.
2015/030	Financing patterns of European SMEs: An Empirical Taxonomy November 2015.
2015/031	SME Securitisation – at a crossroads? December 2015.
2015/032	European Small Business Finance Outlook. December 2015.
2016/033	Evaluating the impact of European microfinance. The foundations. January 2016.
2016/034	The European Venture Capital Landscape: an EIF perspective. Volume I: the impact of EIF on the VC ecosystem. May 2016.
2016/035	European Small Business Finance Outlook. June 2016.
2016/036	The role of cooperative banks and smaller institutions for the financing of SMEs and small midcaps in Europe. July 2016.
2016/037	European Small Business Finance Outlook. December 2016.
2016/038	The European Venture Capital Landscape: an EIF perspective. Volume II: Growth patterns of EIF-backed startups. December 2016.